The University Of Sheffield.

Understanding and Predicting Symptom Trajectories in Psychological Care

Claire E. Bone

University of Sheffield

May 2019

Declaration

This thesis has been submitted as partial fulfilment of the Doctorate in Clinical Psychology qualification. This work has not been submitted for any other degree or to any other institution.

Structure and Word Counts

	Excluding tables, references and appendices	Including tables, references and appendices
Section one: Systematic Review	7,832	12,975
Section two: Empirical Project	8,166	15,908
Total Thesis	15,998	28,883

This page is left intentionally blank

Lay Summary

Understanding and predicting responses to psychological therapy is important to ensure adequate provision of services and to improve outcomes. The first section of this thesis reviews the literature on the "Good-enough level" (GEL) model, which states that people respond at different rates to therapy and "responsively regulate" treatment length. On the surface, this appears to be in conflict with the "Dose-response" model, which states that people respond within defined ranges, following a trajectory of steeper gains in earlier sessions with diminishing effects over time.

Overall review findings supported the GEL. Higher numbers of sessions were not generally associated with more change. Where studies did find links between increased sessions and change, this could be explained by the inclusion of early drop-outs, which creates the illusion of higher doses being more effective. Higher baselines were associated with longer treatments, and different people were found to respond at different rates to therapy, with shorter treatments linked to faster progress.

Findings on the shape of change varied. Some studies linked longer, slower progress with linear trends, and some linked shorter rapid progress with loglinear trends. When different sub-classes of people were examined, however, other differential treatment responses were evident. It appears as though that individual responses to therapy do vary in line with the GEL. However it was also noted that this variation might nonetheless be captured within certain boundaries as described by the dose response literature, where higher treatment lengths do not necessarily lead to better outcomes. Both models could co-exist within a concept of boundaried responsive regulation.

There have been significant developments in the use of technology to understand and predict psychotherapy outcomes. The second section of this thesis details the development of a dynamic progress feedback system. The system was designed to provide personalised prognoses of recovery that update as new information is entered. Statistical modelling techniques combined predictor variables based on profile information from the Leeds Risk Index ([LRI] Delgadillo et al., 2016) with incoming routine progress scores. This included cumulative predictors based on risk spikes and standard deviations, meaning that the models learned from previous information. The models were developed using data from one Improving Access to Psychological Therapies (IAPT) service, and cross-validated in a new IAPT dataset, to assess generalisability. Models of increasing complexity were also created, to examine whether complex models outperformed simpler versions.

Results indicated that for low intensity treatment, models of medium complexity performed as well as the most complex model including the LRI. However at high intensity, the complex model was superior between sessions one to six. All models cross-validated in the new sample apart from a basic LRI profile only model. More complex models tended to see better cross-validation figures, however confidence intervals overlapped suggesting they were not significantly better. Preference for added complexity may therefore depend on service capabilities to implement the LRI. These models are intended for development as a clinical application, where further cross-validation and research to assess their impact on outcomes in practice is intended.

Acknowledgements

I want to thank the participants who shared their outcomes data to make this research possible and I hope it is helpful. I also wholeheartedly thank my research supervisor Dr Jaime Delgadillo for his enthusiasm, clear explanations of complex topics, and for keeping me at the top of my ZPD! I'm grateful to our research collaborators, in particular Dr Richard Thwaites, Dave Sandford and the team at IAPT Cumbria, and Professor Wolfgang Lutz, Professor Julian Rubel and Dr Ann-Kathrin Deisenhofer from the University of Trier for being wonderful hosts. I have also had some other fantastic supervisors over the last 10 years, in particular Michelle, Babak, Nic, Peter, Lou, Jo H, Nick, Fraser, Sue and Jo B.

I have a list of personal thanks: My fellow trainees for the WhatsApp fun and making this experience so memorable (I can't name you all but in particular Lewis Hanney for rating my reliability and his cheery disposition), my Jam dance friends, my old friends (Nursey, Sophie, Spam, Loz), family past (Tom, Harry, John, my Grandmas) and present (particularly my Dad, and Jac, Ocean, Adam, Sonia, Josh, Viki, and my brilliant little sisters Amelia and Lydia). Huge thanks always to my partner Stu who encouraged and supported me to do this from the beginning, and tolerated me colonising our home, and to his parents Pete and Jan for looking after us (and not forgetting Sime and Zoe). Finally I want to thank my Mum, who died when I was 15. She was a Northern Soul loving computer programmer and completed a psychology degree when I was a kid. She gave me the tenacity to work through a stats hole and would have been thrilled to see this.

This page is left intentionally blank

Table of Contents

Section One: Literature Review

Responsive Regulation of Psychotherapy Duration: A Systematic Review and Meta-Analysis of "The Good-Enough Level" Literature

Section Two: Empirical Study

The Development of a Dynamic Progress Feedback System to Guide Psychological Treatment in Primary Care

Section One: Systematic Review and Meta-Analysis

Title: RESPONSIVE REGULATION OF PSYCHOTHERAPY DURATION: A SYSTEMATIC REVIEW AND META-ANALYSIS OF THE "GOOD-ENOUGH LEVEL" LITERATURE

Short title: SYSTEMATIC REVIEW OF THE GOOD-ENOUGH LEVEL

Claire E. Bone

University of Sheffield, Clinical Psychology Unit, Cathedral Court, Floor F, 1, Vicar Lane, Sheffield. S1 2LT.

This page is left intentionally blank

Abstract

Objectives

This review aimed to examine the "Good-Enough Level" (GEL) concept that people respond differently to therapy, as well as examining the shape of change (Barkham et al., 1996).

Methods

Systematic searches took place using Medline, PsycINFO and Scopus databases. Key search terms were variants of: Good-enough level, dose-response, treatment duration, rate of change, treatment outcome, responsive regulation and psychotherapy. A key inclusion criterion was that cases must be stratified by treatment length to examine the GEL. A narrative synthesis was provided, with random effects meta-analysis where possible.

Results

Fifteen studies were synthesised (n=114,123), with five used in primary meta-analyses (n=46,921), and sub-group analyses performed on differential findings. High heterogeneity was observed making conclusions tentative, however there was no overall association between improvement and total sessions, which supports the GEL (r=-0.24 [95% CI -0.70, 0.36], p=0.2747). Increases in improvement associated with longer treatment may be an artefact of people terminating therapy early. Longer treatments were associated with higher baselines (r=0.15 [95% CI 0.08, 0.22], p<.001) and slower rates of change, in support of the GEL. Shapes of change varied, but emerging patterns suggested that longer treatments may see more linear trajectories, challenging the universal curvilinear trend suggested in the dose-response literature.

Conclusions

Support was found for the GEL: treatment length appears to be responsively regulated based on need, and there is heterogeneity in trajectories of change. However this may also occur within boundaries suggested by the dose-response literature. The models could co-exist within a concept of "boundaried responsive regulation".

Keywords: "good-enough level" "psychotherapy outcomes" "dose response" "dose effect" "psychotherapy treatment duration" "responsive regulation" "rate of change"

Practitioner Points

- Services should ideally be planned such that they can responsively regulate treatment, within clinical boundaries suggested by the dose-response literature
- Stratified care is important for future planning of services
- Better recording of treatment endings and client and therapy characteristics would enable future research

Key Limitations

- High heterogeneity and small study numbers meant that quantitative analyses were unable to draw strong conclusions
- The review was limited by similar issues as its component studies, where missing patient and intervention characteristics, and missing outcomes data influence findings and interpretations

Responsive Regulation of Psychotherapy Duration: A Systematic Review and Meta-Analysis of the "Good-Enough Level" Literature

People experiencing mental health problems are commonly offered psychological therapies, however there is debate as to how many sessions are needed in order to optimise treatment response and use of resources. A key goal of patient focused research has therefore been to identify the optimum number of sessions required to see 'improvement'. Improvement is defined in many different ways in the literature, but in UK psychological services and for the purpose of this paper it will be quantified by the concept of reliable and clinically significant improvement (RCSI), where symptom severity scores have shown change that statistically reliable (not merely explained by measurement error) and that have also moved to below clinical cut-off levels (Jacobson & Truax, 1991).

The Dose-Response and Good-Enough Level Models

The literature describes two perspectives on the number of sessions required to benefit from therapy: 'Dose-response' models (DR) and "Good-enough level' models (GEL). Dose-response models initially arose in the medical literature and describe a curvilinear (negatively accelerating curve [NAC]) shape of change. Howard, Kopta, Krause and Orlinksy (1986) first applied the concept to psychotherapy settings, performing a meta-analysis to examine how many sessions were needed to see 50% or 75% improvement. They also found the curvilinear pattern of response, which has been instrumental in helping services plan treatment lengths and maximise resources.

The interpretation of the NAC is that it represents the average of multiple individual curvilinear responses. Response to treatment is considered to be initially steep before steadily reducing as therapy progresses. Accordingly, the amount or 'dose'

of therapy influences improvement, and there is a pattern of diminishing improvements with each subsequent treatment session. The assumption is that most people tend to follow this pattern and therefore a generalised optimum amount of sessions can be identified. In line with this, the probability of recovery is assumed to be correlated with treatment length (Howard et al., 1986). This model has been highly influential for the development of symptom monitoring and feedback systems that use 'expected treatment response' (ETR) curves. These model expected trajectories of symptomatic change and allow clinicians to observe when clients are "not on track" to see improvement (Lutz, Martinovich, & Howard, 1999).

The GEL perspective has attempted to expand on this by further examination of sub-groups of patients. For example, Barkham et al. (1996) pointed out that the majority of psychotherapy research on the dose-response examines different samples at different time-points, since different patients complete therapy following different 'doses'. This effect of aggregating across different groups has an important implication for how results are interpreted. For example, some people may terminate therapy after few sessions due to having steep early progress, whereas others have slower progress and require longer support. The decelerating shape of change may therefore be an artefact of rapid improvers in early samples and slower to respond cases in remaining samples, until there may be only non-responders left (who remain in treatment nonetheless). Indeed there is now a growing body of evidence examining differential patterns of response to psychotherapy (Delgadillo, McMillan, Lucock, Leach, Ali, & Gilbody (2014); Lutz et al., 2012; Rubel, Lutz, & Schulte, 2015).

Therapy length in routine care has therefore been argued to result from 'responsive regulation' by clients and clinicians (Stiles, Honos-Webb, & Surko, 1998). Rather than length of therapy determining progress (e.g. the number of sessions *causing*

improvement), the GEL perspective states that progress determines therapy length, with treatment ending when improvement reaches a 'good-enough level'. The rate of change therefore varies depending on treatment length or dose, whereas average rates of recovery do not. According to this model, slopes (or rates) of change vary for different people (for various reasons including nature of difficulty, life circumstances, therapist effects, etc. see Barkham et al., 1996; Goldberg, Hoyt, Nissen-Lie, Nielsen, & Wampold, 2018). The probability of recovery would be considered to be unrelated, or negatively related, to treatment length, where slow or non-responders have a reduced chance of recovery (Barkham et al., 2006).

Dose-effect curves are taken from medicine, where too low a dose produces little effect, then there is a 'therapeutic window' where medication produces good effect, followed by too high a dose, where benefit declines into potential harm. Conversely, in agriculture, the dose-effect has been understood as some organisms being more resistant to poisons than others. For example, initially a poison may have rapid effect on killing insects however this tapers off until more resistant insects are left. Here the strength of the poison is the same, but the effect is less, which is analogous to the GEL (Barkham et al., 2006). This represents a population view, where fewer people are likely to benefit from low doses or need high doses, with the majority finding a GEL within mid-ranges (Baldwin, Berkeljon, Atkins, Olsen, & Nielsen, 2009).

This is an important matter given the implications for the wellbeing of clients, and the debate between the two models has led to a body of research on the topic, including two informative book chapters (Castonguay, Barkham, Lutz, & McAleavey, 2013; Nielsen, Bailey, Nielsen, & Pedersen, 2016). Key differences between the two perspectives can be summarised as follows:

Table 1

Key Differences Between DR and GEL models

Dose-Response	**GEL**
Curvilinear response is an average of multiple individual curvilinear responses	Curvilinear response is an artefact of aggregating people, where faster remitters end therapy earlier
Rate of change does not vary with total sessions	Rate of change does vary with total sessions
Improvement is associated with total sessions	Improvement is not associated (or negatively) with total sessions
Therapy length determines progress	Progress determines therapy length

Rationale for a Systematic Review

A recent systematic review examined 26 papers on the dose-response effect in psychotherapy, finding overall support for a curvilinear response to treatment. However, this review also included six studies that contrasted DR models with GEL models, where unanimous support was found for the GEL proposal that different people respond to therapy at different rates (Robinson, Delgadillo & Kellett, 2019).

An important limitation of this prior review is that it is likely to have excluded many relevant GEL-oriented studies, since the focus of that review was specifically on the DR model. There are currently no existing systematic reviews on the GEL literature. If the GEL interpretation is empirically supported, this could have important implications for how far prototypical responses to treatment can be generalised. It could impact on how services are organised and commissioned, and inform interpretations of psychotherapy outcome research. For example, policies are formed based on optimum doses, and theories are informed on how clients change (e.g. if people on average see

diminishing gains, or if on average they make linear progress, this changes how the clinician interprets progress).

Current outcome monitoring and feedback models that guide treatment are largely based on ETRs, which are informed by current dose-response research (Lambert & Shimokawa, 2011). However given evidence that not all people may follow a NAC, methods of monitoring outcomes are evolving. Section two of this report details the development of a dynamic progress feedback system, which attempts to capture more of this individual variance.

Aims

A systematic review of the GEL literature was therefore carried out. The aim was to synthesise findings that stratify psychotherapy treatment by sub-groups having the same treatment length, to understand whether different people respond to therapy at different rates, and whether therapy tends to terminate when people have reached a GEL. Although the GEL model does not specify a shape of change, it was also of interest to understand whether the shape of change within sub-groups is linear or non-linear, since the curvilinear shape has been proposed as a key tenet of the DR model. The PICO framework was used to develop the review questions and to identify inclusion and exclusion criteria as described in table 2 (Cherry & Dickson, 2014). The following research questions were posed:

i. Do different sub-groups of adults accessing psychotherapy respond to treatment at different rates in line with the good-enough level perspective?

ii. Is the shape of change linear or non-linear?

Method

Protocol Registration

The review protocol was published prospectively on the PROSPERO register at https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=131840.

Search Strategy

Medline (OvidSP), PsycINFO (OvidSP) and Scopus databases were searched by title, abstract, keywords or subject headings, with no date restrictions. Further searches were performed using Google Scholar and forwards and backwards reference list searches, and all identified GEL study authors were contacted to check for missed papers. Search terms were combined using Boolean operators (AND / OR) and truncation or wildcards (e.g. *, ?). Keywords were variants of: Good-enough level, dose-response, treatment duration, rate of change, treatment outcome, responsive regulation and psychotherapy (appendix A).

Although the focus was on the GEL, the term "dose-response" was included to capture any studies that contrasted the two models. A pragmatic decision to exclude grey literature was made, due to the unlikelihood of finding further quality research in this field given the contact with experts, and the importance of peer review given the methodological complexity of the topic. Some services include people aged 16 within adult psychotherapy settings, therefore 16 was used to define adulthood.

Inclusion and exclusion criteria. Table 2 describes the inclusion/exclusion criteria and review questions framed by PICO domains. *N*=2299 records were initially identified. Ten of the GEL authors responded, with one providing an extra chapter containing primary research. This left *n*= 2083 after removing duplicates. Titles and abstracts were screened for relevance, followed by full-text reviews. Following

screening, *n*=15 papers were included in the qualitative synthesis as described in the PRISMA diagram below.

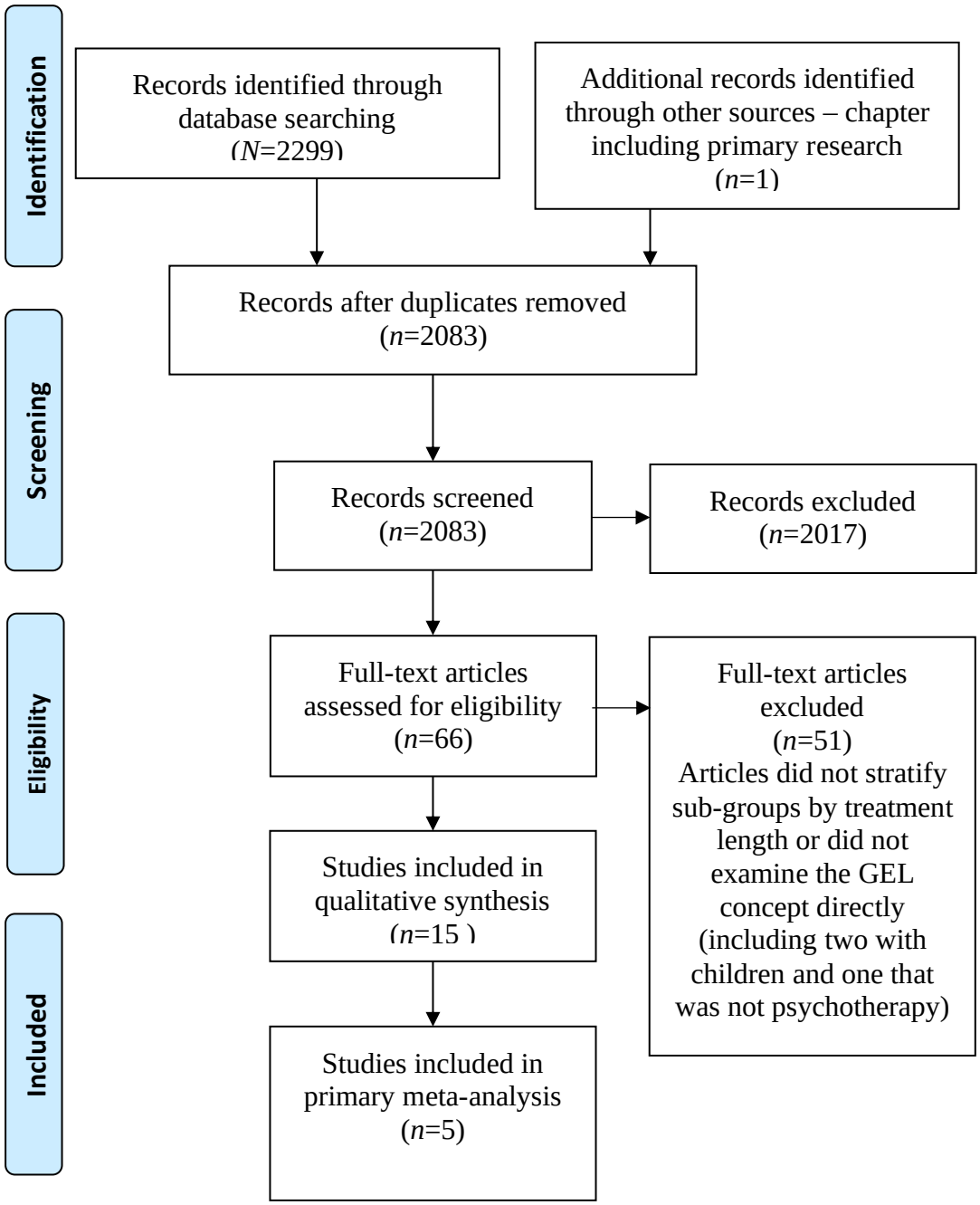


Figure 1. Prisma diagram based on Moher, Liberati, Tetzlaff, and Altman, 2009.

Table 2

Review Questions and Inclusion and Exclusion Criteria

	Review questions	
	Do different sub-groups of adults accessing psychotherapy respond to treatment at different rates in line with the "Good enough level" perspective? Is the shape of change linear or non-linear?	
	Inclusion criteria	**Exclusion criteria**
Population	People over 16 accessing psychotherapy treatment.	Studies researching children and/or adolescents under 16.
Intervention	Any form of psychological intervention, delivered in any format.	Studies that do not include psychological interventions.
Comparator	Study design must stratify cases by treatment length and examine associations between treatment duration and outcomes based on the GEL concept directly. Studies may also compare different shapes of change.	Studies where cases are not compared by treatment length, for example only examining aggregate group responses to identify rates and shape of change.
Outcomes	Response to psychotherapy 'dose' measured using standardised outcome measures, examining the rates and/or shape of change.	Studies that do not use standardised outcome measures or measure outcomes as a result of non-psychological interventions. Studies that do not examine either rate or shape of change in response to psychotherapy.
Setting	Any settings where psychological interventions are usually delivered, across clinical and non-clinical settings (including outpatient, inpatient, university counselling centres, etc.), in any country.	Non-psychological intervention settings.
Study design	Practice-based naturalistic studies or controlled trials of psychological interventions. Cases must be stratified by treatment length. Studies published in English in peer reviewed scientific journals.	Studies that do not use a stratified design (by treatment length). Literature not published in peer reviewed scientific journals. Research studies not in published in English.

Data Extraction

Data extraction included the following elements: aims, setting, sample number, demographics, inclusion/exclusion criteria, presenting problem, intervention, outcome measures, outcome criteria, design, method, analysis, treatment duration, findings, effect sizes, and conclusions (appendix B).

Risk of Bias Assessment

Risk of bias was assessed using a customised tool based on the Critical Appraisal Skills Programme Cohort Study Checklist ([CASP] 2018) and Cochrane guidance (Higgins & Green, 2011) (appendix B). Cochrane and CASP do not recommend using an overall scoring system and prefer the use of 'yes, no, unclear' criteria. Ratings were completed independently by two reviewers on all papers, and Cohen's Kappa used to assess inter-rater reliability. Disagreements were discussed and resolved between the two raters.

Data Analysis

Findings were synthesised in narrative format, structured by the different methodological approaches to examining the GEL. Random-effects meta-analysis was also performed where data permitted, using the R package Meta-Analysis via Shiny ([MAVIS; Version 1.1.3] Hamilton, 2017). Heterogeneity was examined using Cochrane's Q and the I^2 statistics (Higgins & Thompson, 2002). High heterogeneity was observed, therefore subgroup analyses were also performed to examine potential methodological sources of heterogeneity a priori. Moderator regression analyses were not possible due to insufficient and imbalanced study numbers (Borenstein, Hedges, Higgins, & Rothstein, 2009). There is debate as to whether small study numbers should be used in sub-group analyses, or meta-analyses more generally. However as Borenstein et al. (2009) point out, people tend to draw conclusions whether this is done narratively

or statistically thus it is better to do this in an informed way. All analyses are therefore drawn together to inform discussion in the review.

Publication Bias

Publication bias refers to the problem of non-significant findings tending not to be published, leading to a biased representation of findings on a topic (Field & Gillett, 2010). Formal tests to examine publication bias were applied, including the regression test for funnel plot asymmetry (Egger, Smith, Schneider, & Minder, 1997), and the weight-function likelihood ratio test (Vevea & Hedges, 1995).

Results

Study Characteristics

Fifteen studies were included in the final review (table 5). Of these, 14 were database analyses of naturalistic psychotherapy outcomes data, with one using random allocation to fixed treatment lengths (Barkham et al., 1996). Gottfredson, Bauer, Baldwin and Okiishi (2014) provided a re-analysis of data from Baldwin et al. (2009), therefore their data only used for discussion. Stiles, Barkham and Wheeler (2015) report that there may be up to 1.8% data overlap between their study and Stiles, Barkham, Connell and Mellor-Clark (2008), and Barkham et al. (2006). There was also database overlap between Owen et al. (2015) and Owen, Adelson, Budge, Kopta and Reese (2016), however they examine different aspects of the GEL in each paper and are not treated as unique samples for aggregation.

Sample and setting. The total sample across studies was N=204901, with n= 114123 included in the main GEL analyses (excluding Gottfredson et al. 2014, and acknowledging some overlap). Five studies were UK-based (mixed settings), nine were US-based (all university counselling centres apart from one community centre), and one from Sweden (primary and psychiatric samples).

Measures. Six outcome measures were used: The Beck Depression Inventory (BDI), the Inventory of Interpersonal Problems (IPP-32), and the Personal Questionnaire (PQ) were used in Barkham et al. (1996); the Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM) was used in five studies; the Outcome Questionnaire-45 (OQ-45) in five studies; and the Behavioral Health Measure (BHM) in four. The BDI, IPP and PQ are specific measures of depression, interpersonal problems and idiographic items respectively. However, the remaining measures assess general psychological distress.

Outcome criteria. The criteria for change were described differently across studies, however they all referred to either the concept of reliable change (RC) or reliable and clinically significant improvement (RCSI), as defined by Jacobson and Truax (1991) and discussed in the introduction.

Interventions. A wide variety of interventions were used under the psychotherapy domain, including integrative approaches using more than one approach in combination. Several studies had limited information on interventions used.

Risk of Bias Assessment

All of the studies were considered relatively low risk of bias (appendix B). However there were common aspects of methodological bias or variance inherent in naturalistic databases, relating to: missing data, planned versus unplanned endings, missing ending information, missing demographic information, missing intervention information, varying outcome criteria, and varying modelling strategies. These issues will be discussed alongside findings in the relevant sections.

Cohen's Kappa found moderate agreement between raters k=.507, p<.001 (Altman, 1999), where ratings matched 85% of the time. On discussion, the majority of disagreements were on whether authors had identified and overcome all confounds (yes

versus unclear) and whether there were missing data that were unaccounted for (one rater marked unclear impact of bias if methods were used to handle missing data or discussion of impact, whereas the other considered this still to be missing data). It was agreed that for studies where clear discussions of missing data were provided this would be classed as 'unclear risk', where a distinction is made between them and other studies not addressing the issue.

Methods of Examining the GEL

Four approaches to examining the GEL were used: (i) associations between improvement and total sessions, (ii) associations between baseline symptom scores and total sessions, (iii) assessing rates of change, (iv) assessing the shape of change.

Table 3

Summary of Approaches and Methods of Examining the GEL

Approach	Method
Assess whether improvement is associated with total sessions.	Correlate improvement (RCSI or RC) with total sessions. Calculate % seeing RCSI or RC and report descriptively whether rates appear consistent. Use structural equation modelling to assess whether sessions predict change, or change predicts sessions.
Assess whether different people require different amounts of therapy.	Correlate baseline symptom severity with RCSI or RC, or use baseline as a regression predictor. Create models using patient and therapist effects.
Assess whether rate of change varies with total sessions.	Use modelling techniques to model individual change, or generate groups of clients with similar profiles of change. Correlate mean rate of individual change with total sessions. Incorporate session frequency variables.

Assess the shape of change.	Use modelling techniques to assess fit of linear, log-linear, cubic, quadratic, or other shapes of change.

Table 4

Summary of Modelling Techniques and Abbreviations Used

Abbreviation	Description
MGCM	Multi-level growth curve model (Broad term which examines between groups comparisons of intra-individual change)
MLM	Multi-level model (Can include MGCM – allows for nesting of data at different levels)
LGCM	Latent growth curve model (See MGCM)
MLGM	Multi-level growth model (See MGCM)
GMM	Growth mixture model (See MGCM – particularly regarding the identification of sub-classes based on similarities)
SEM	Structural equation modelling (allows examination of predictive relationships between latent or unobservable variables)
SPMM	Shared parameter mixture model (allows for the modelling of missing data when reasons are not known)

Notes. See Curran, Obeidat & Losardo, (2010); Field (2018); Gottfredson et al. (2014).

Table 5

Study Characteristics

First Author and Year	Study Design	Study Setting	Presenting Problems	Total *N* (204,901)	Analysed *n* (114,123)	Intervention	Outcome Measures/criteria	Duration
1. Baldwin et al. (2009)	Database analysis	US University counselling centre	Mixed	4676	2985 above cut-off	Mixed	OQ-45 RCSI	Mean 6.46 sessions
2. Barkham et al. (1996)	Random allocation	UK Psychotherapy settings	Mixed, with 85% depression	212	106 in 8 105 in 16	CBT or PI	BDI, IPP-32, PQ	Fixed, 8 or 16 sessions
3. Barkham et al. (2006)	Database analysis	33 UK NHS Primary care	Mixed	1868	1472 above cut-off	Mixed	CORE-OM RCSI/RC	Some fixed but flexible, PE, 12 sessions or less
4. Erekson et al. (2015)	Database analysis	US University counselling	Mixed	22,235	21488	Mixed	OQ-45 RCSI	Mean 5.8 sessions
5. Evans et al. (2017)	Database analysis	UK Secondary care	Mixed	4877	925	Mixed	CORE-OM RC	Median 15 sessions, 26 weeks, .61 per week
6. Falkenström et al. (2016)	Database analysis	Swedish Primary and psychiatric services	Mixed	1794	924	Mixed	CORE-OM Scores modelled	Mean 6 primary care / 9.1 psychiatric
7. Gottfredson et al. (2014)	Database re-analysis	US University counselling	Mixed	4676	2985	Unknown	OQ-45 Scores modelled	Median 8 sessions/6.89 weeks

		(Baldwin et al. 2009)							
8.	Kivlighan et al. (2018)	Database analysis	US University counselling	Unknown	786	438 / 369 with ending info	Unknown	BHM-20 Scores modelled	Some PE. Mean 5.54 sessions
9.	Nielsen et al. (2016)	Database analysis	US University counselling	Mixed	24,860	17,490	77.8% individual, then mixed.	OQ-45 RC	Median 4, modal 1 (1-548)
10.	Owen et al. (2015)	Database analysis	47 US College counselling centres & 1 community centre	Unknown	38,985	10,854	Unknown	BHM Scores modelled	Mean 9.41, median 8 sessions
11.	Owen et al. (2016)	Database analysis	46 US College counselling centres & 1 community centre	Unknown	48,963	13,664	Unknown	BHM RC / scores modelled	Mean 9.04 sessions
12.	Reese et al. (2011)	Database analysis	US University counselling	Mixed	3270	1207	Mixed	OQ-45 Scores modelled	90% <15 sessions, median 5
13.	Stiles et al. (2008)	Database analysis	UK 32 Primary care services	Mixed	9703	9703	Mixed	CORE-OM RCSI / mean change	PE, <=20 sessions. Some fixed=6 but flexible
14.	Stiles et al. (2015)	Database analysis	UK NHS 6 Primary care, 8 secondary care, 2 tertiary care, 10	Mixed	36,297	26,430	Mixed	CORE-OM RCSI	PE, Some fixed (6) but flexible, median 6 sessions.

			University, 14 voluntary, 2 private						
15.	Stulz et al. (2013)	Database analysis	US 20 College counselling centres, 4 primary care centres, 2 private centres.	Mixed	6375	6331	Mixed	BHM RCSI	Median 5 sessions

Notes. All endings include unplanned unless marked 'PE' for planned endings. US: United states, UK: United Kingdom.
Outcome measures abbreviations: BDI: Beck Depression Inventory (BDI); IPP-32: Inventory of Interpersonal Problems-32; PQ: Personal Questionnaire; CORE-OM: Clinical Outcomes in Routine Evaluation – Outcome Measure; OQ-45: Outcome Questionnaire-45 (OQ-45); BHM: Behavioral Health Measure.

Table 6

Findings Reported by Approach and Method Used

First Author and Year	Method	Reported Findings/Statistics
	Associations between improvement and total sessions	
Baldwin et al. (2009)	Logistic regression using total sessions as predictor of RCSI. Min=3 sessions. RCSI binary. Correlation between sessions totals and final scores.	Small non-linear relationship between RCSI and total sessions – small increase up to session 8, then rates of RCSI plateau. Loglinear term significant for sessions and RCSI, odds ratio: 3.08, $p<.05$ Converted to $r=0.2962$ for meta-analysis. No correlation between sessions and final scores $r=.02$, $p=.09$.
Barkham et al. (2006)	Percentage calculation of RCSI per group. Correlation between rate of RCSI and total sessions	Large negative correlation between rates of RCSI and total sessions $r=-.91$, $p<.001$ (up to 12 sessions)
Evans et al. (2017)	Correlation between change in score and total sessions. Min=3 sessions. Examined	No correlation between change in score and total sessions $r_s= -.04$, $p=.289$.

	differences between reliable change categories and dose.	No significant differences between reliable change groups and total sessions, $H(3)=.67$, $p=.879$
Owen et al. (2016)	Regression between amount of change on items and total sessions	Small associations on individual items: Wellbeing: $r_2=.014$ Symptom distress: $r_2=.021$ Life functioning: $r_2=.004$
Nielsen et al. (2016)	Linear correlation between change and total sessions. Linear and non-linear regressions using various terms between change scores and total sessions. SEMs to analyse regressions of symptom change on sessions (sessions predict change - DR) and sessions on change (change predicts sessions - GEL). Plus a combined DR and GEL SEM. Analysed with X_2	No linear correlation $r=.008$, p=.29. However inverse (NAC) regression significant: $F(1, 17488)=72.5$, $p<.001$, $R_2=.004$. Increases in change seen up to session 18 then plateaus. When reliable change criteria is used, plateau occurs at 6 sessions. SEMs showed that the only adequate fit was achieved by a DR *plus* GEL SEM: $X_2(1, n=17490)=2.5$, $p=.065$. Variance explained was improved by individual therapy modality effects (.02% to 13%).
Stiles et al. (2008)	Percentage calculation of RCSI per group Correlation between RCSI / RC and total sessions. Compare mean pre-post change scores by total sessions.	Change scores similar across treatment lengths Large negative correlation between RCSI and total sessions No correlation between RC and total sessions RCSI: $r= -.75$, p<.001 RC: $r=.11$, ns
Stiles et al. (2015)	Percentage calculation of RCSI per group Correlation between rates of RCSI / RC and total sessions. Compare mean pre-post change scores by total sessions.	Change scores similar across treatment lengths. Large negative correlation between RCSI and total sessions Moderate negative correlation between RC and total sessions RCSI: $r=-.58$, p<.001 RC: $r=-.40$, p<.001.
Stulz et al. (2013)	Correlation between rates of RCSI and total sessions. Min=3 sessions.	Large positive correlation between RCSI and total sessions $r=.714$, p=.004

Associations between baseline symptom scores and total sessions

Baldwin et al. (2009)	Correlation between baseline score and total sessions. Min=3 sessions.	Small positive correlation between baseline score and total sessions. r= .09, p<.001
Barkham et al. (2006)	Correlation between baseline score and total sessions	Small positive correlation between baseline score and total sessions. r= .13, p<.001
Erekson et al. (2015)	MLM with linear, quadratic and cubic terms. Min=2 sessions.	Higher levels of dose associated with lower levels of OQ-45 at intercept.
Evans et al. (2017)	Correlation between baseline score and total sessions. Min=3 sessions.	Small-moderate positive correlation between baseline score and total sessions. r= .29, p<.005
Falkenström et al. (2016)	MLGMs comparing DR and GEL models to assess whether rate of change varies as function of treatment length. Min=3 sessions.	Although they found that initial symptom severity was not related to treatment length in weeks, the psychiatric sample had higher risk and higher total sessions numbers.
Owen et al. (2015)	3-level model, initial scores nested in clients nested in therapists. Min=4 sessions.	Clients in different classes showed differences in intake scores – 'Early & Late', and 'Slow & Steady', had higher intake scores than 'Worse Before Better'. Slow & Steady more distressed and slower trajectory overall.
Stiles et al. (2008)	Correlation between baseline score and total sessions. Correlation between mean baseline scores and total sessions.	Small positive correlation between baseline score and total sessions. r=.16, p<.001 Large positive correlation between mean baseline score and total sessions r=.93, p<.001.
Stiles et al. (2015)	Correlation between baseline score and total sessions. Correlation between mean baseline scores and total sessions.	Small positive correlation between baseline score and total sessions. r=.08, p<.001 Large positive correlation between mean baseline score and total sessions r=.58, p<.001.

	Assessing rates of change	
Baldwin et al. (2009)	MGCM – compared average rate of change with total sessions. Min=3 sessions.	Significant interaction between rate and dose, slower rates associated with higher dose. Log of total sessions and cubic form: cubic (beta): 0.02, $p<.01$. Interactions between log of total sessions and time: Linear =2.69, Quad= -.29, Cubic=.02, all p<.01.
Barkham et al. (1996)	Percentage calculation of RCSI per group (8 or 16 sessions).	8 session group had faster rates of improvement than 16 session group at 8 sessions on BDI ($X^2(1, n=181)=6.03$, $p=.014$) and PQ items. However not on IPP-32. On BDI – faster reductions in distress, slower in characterological/interpersonal. Explains slower rates on IPP, also seen in PQ items.
Erekson et al. (2015)	MLM with total sessions and session frequency as continuous variable on rate of change. Min=2 sessions.	Higher doses had slower improvement rates, less frequent sessions had slower rates of change. Adding session frequency improved BIC by 8,515. Rate of change (based on clinically significant change) was faster in weekly than fortnightly groups based on total sessions: $X^2= 39.36(1)$, $p<.001$. Effect size of session frequency f^2 0.07.
Falkenström et al. (2016)	MLGMs comparing DR and GEL models to assess whether rate of change varies as function of treatment length. Min=3 sessions.	GEL model a better fit in primary ($X^2(4) = 37.46$, p<.001) and psychiatric ($X^2(3) =25.68$, p<.001) samples. Faster rates of change with fewer sessions in both samples, but psychiatric saw slower rates of change and higher total sessions.
Gottfredson et al. (2014)	SPMMs used to re-analyse data from Baldwin et al. (2009), to handle missing data.	SPMMs indicated that faster responders were more likely to terminate therapy earlier, meaning rates of change underestimated (6.50% - 6.66% across two models).

Kivlighan et al. (2018)	MLM estimated with linear, log-linear and quadratic terms – measure broken down into different domains and dependency between items controlled for. Min=2 sessions. Analysed planned vs unspecified endings.	Log-linear best fit for all >=2 sessions, linear best fit for all >=3 sessions. Rate of change did not vary on individual domains, but did overall: (-0.01, p = .024). People more likely to terminate early due to changes in wellbeing but not other items.
Owen et al. (2015)	GMM. Identified 3 different classes (1. Early and late, 2. worse before better, 3. slow and steady). Modelled linear, quadratic and cubic rates of change. Min=4 sessions.	All were significant, initial rates of change (over first 3 sessions) differed – Slow and steady class had slower rate of change than early and late, and worse before better. Coefficients on initial rates of change: Slope Class 3 vs Class 2: 22.75, Class 1 vs Class 3: 4.93, p<.001.
Owen et al. (2016)	MLMs estimated rate of change for DR and GEL models and compared fit. Min=1 session. On individual questionnaire domains.	GEL Log-linear model was best fit for wellbeing and symptom distress (Loglinear x sessions interaction coefficients: -0.0098 / -0.0081, p<.01). GEL quadratic model best fit for life functioning ($Session^2$ x sessions interaction coefficient: 0.0002, p<.01). Clients attending fewer sessions had faster rates of change. However change on life functioning was smaller than wellbeing or symptom distress. Therapist effects explained some of variations in change on wellbeing and life functioning.
Reese et al. (2011)	MLGM with improvement as a function of total sessions and session frequency. Used linear, cubic and quadratic terms.	GEL model significantly better fit than DR, longer sessions had slower rates of change. GEL modified (including session frequency) was significantly better fit than GEL alone, less frequent sessions had slower rates of change. GEL: X2(2)=98.2, p<.001. GEL vs GEL mod:X2=18.1, p<.001. Overall linear trends most parsimonious – linear and steeper at <5.72 sessions.
Stulz et al. (2013)	LGCMs – correlated mean rates of change with total sessions. Min=3 sessions.	Large negative correlation between mean change and total sessions: r= -.974 (for log-linear model – best fit).

	Assessing shape of change	
Baldwin et al. (2009)	MGCMs compared DR and GEL, modelled as linear based on previous studies then cubic based on visual inspection. Measures every session. Min=3 sessions.	DR model produced NAC, however GEL model fit with cubic terms superior (double curve) $X_2(4)$=428.49, p<.0, Cubic beta= -.06, p<.01. Cubic BIC: 244,425
Barkham et al. (1996)	Percentage calculation of RCSI per group Pre, mid (for 16 sessions), and post therapy.	Linear improvement seen on PQ items and in sequence of RCSI percentages on BDI or IPP. When aggregated across both groups however Log-linear NAC shape seen.
Erekson et al. (2015)	MLM with linear, quadratic and cubic terms. Min=2 sessions.	All significant but linear largest estimate.
Falkenström et al. (2016)	MGLMs comparing DR and GEL models using linear, quadratic and cubic terms. Min=3 sessions.	GEL model a better fit in primary ($X_2(4)$ = 37.46, p<.001) and psychiatric ($X_2(3)$ =25.68, p<.001) samples. In primary care: Linear, cubic and quadratic all significant but quadratic shape best. In psychiatric sample linear shape best.
Kivlighan et al. (2018)	MLMs estimated with linear, log-linear and quadratic terms – measure broken down into different domains and dependency between items controlled for. Min=2 sessions.	Log-linear best fit for all >=2 sessions (BIC 35,728.83), linear best fit for all >=3 sessions (BIC 3320.65).
Nielsen et al. (2016)	Linear and non-linear terms used in regression analyses of change scores and total sessions. Then used SEM to identify more complex relationships between shape of change and whether total sessions predict improvement or improvement predicts total sessions.	Inverse (NAC) regression significant/largest: $F(1, 17488)=72.5$, $p<.001$, $R^2=.004$. Increases in change seen up to session 18 then plateaus. When criteria of reliable change is used, rates plateaued by the 6th session. Higher sessions fit GEL, shorter fit DR. Combined DR and GEL SEMs fit data best.

Owen et al. (2015)	GMM to identify sub-classes. Modelled linear, quadratic and cubic forms. Min=4 sessions.	3 classes model significant: Class 1 = early and late change (largest), Class 2=worse before better (smallest), Class 3=slow and steady (linear, longer therapy). AIC: 1, 087, 760. Adjusted BIC: 1, 087, 957.
Owen et al. (2016)	MLMs – Compared fit for log-linear, cubic and quadratic terms for DR and GEL models. On individual questionnaire domains. Measures every session. Min=1 sessions.	GEL better fit than DR. GEL Log-linear model was best fit for wellbeing and symptom distress, quadratic on life functioning. Clients having fewer sessions saw log-linear trend, those having longer sessions saw more linear trend. Wellbeing: GEL Log-linear BIC: 201, 622. Symptom distress: GEL Log-linear BIC:121,483. Functioning: GEL quadratic BIC: 174,939
Reese et al. (2011)	MLGMs - compared aggregate, GEL, and GEL with session frequency. Used linear, cubic and quadratic terms. Measures every third session. Min=1 session.	GEL with session frequency best fit. The GEL model also explained 3% more variance in scores than DR. Cubic terms significant but non-linear trend very subtle so linear terms used. GEL vs GEL modified: $X_2(2)=18.1$, $p<.001$ GEL modified AIC=30, 709.4. Overall linear trends most parsimonious – linear and steeper at <5.72 sessions.
Stulz et al. (2013)	LGCMs – compared linear and log-linear stratified models. Min=3 sessions. Measures every session	Log-linear outperformed linear regardless of treatment length. (Online supplement figures not available).

Notes. Where studies refer to comparisons between the DR model and the GEL model, they mean aggregated or stratified by total sessions received. Min.=3 for e.g., refers to minimum number of sessions.
Model abbreviations: MGCM: Multi-level growth curve model. MLM: Multi-level model. LGCM: Latent growth curve model. MLGM: Multi-level growth model. GMM: Growth mixture model. SEM: Structural equation modelling.

Narrative Synthesis

i) Associations between improvement and total sessions

The GEL model assumes that if people terminate therapy when they have reached a good-enough level, there should be no - or a negative - correlation between total sessions and rates of improvement. Eight studies examined this relationship, with six using correlation analyses and two using regression. Nielsen et al. (2016) also used structural equation modelling (SEM) to expand on their findings.

Five studies found support for the GEL, with no – or negative – correlations between improvement and total sessions. Two of these studies also compared mean change scores by total sessions, finding similar change scores regardless of treatment duration. Two studies concluded partial support, as they only evidenced small associations between improvement and total sessions. Stulz, Lutz, Kopta, Minami, & Saunders (2013) however found a large positive correlation between RCSI and total sessions.

Although Nielsen et al. (2016) found no linear correlation between change and total sessions, they followed this up using regression and SEMs. Regression analyses indicated that the amount of improvement did increase up to session 18 then plateaued. When the criteria of reliable change was used, the plateau occurred at session six. They further analysed this using SEMs, explaining that although studies have examined the relationship between dose and outcome, none have yet examined the direction. SEMs indicated that the only adequate fit for the data was a combined DR and GEL model (and note that including individual therapy as a modality improved on the variance explained from .02% to 13%). In other words, sessions could predict change, but only in a model where it was also possible for change to predict sessions.

There appear to be various possible factors contributing to different results in these studies and the findings are explored further using meta-analysis below.

ii) Associations between baseline symptom scores and total sessions

The GEL model states that different people require different 'doses' of therapy. Eight studies examined relationships between initial symptoms scores and total sessions. Five of these correlated baselines with total sessions, finding small or moderate positive correlations suggesting that people with higher baseline scores have longer treatment. Two of these studies also correlated mean baseline scores with total sessions, finding large positive correlations. In addition, Owen et al. (2015) used GMM to show that higher baselines were associated with different sub-classes of clients, in particular those making 'early and late changes', or 'slow and steady' progress. Falkenström, Josefsson, Berggren, & Holmqvist (2016) used MGLMs to compare primary and psychiatric samples. The psychiatric sample had higher risk and higher total sessions numbers. However note that Erekson, Lambert and Eggett (2015) found that higher total sessions were associated with lower OQ-45 scores at intercept, although in line with the GEL this was associated with longer treatment durations.

Further meta-analysis was performed on those five studies providing correlation coefficients (below).

iii) Assessing rates of change

A further GEL assumption is that rates of change will differ across people accessing different treatment 'doses', with faster remitters leaving therapy earlier. Nine studies assessed whether rates of change differ depending on treatment length. Eight used various modelling techniques to examine this, with one using descriptive methods of comparing % RCSI rates across different questionnaire domains at different

treatment lengths. All nine studies observed that rates of change were faster in those who had fewer sessions. Two studies expanded on this through showing that those having more frequent sessions had faster rates of change, and two found that problems relating to characterological, interpersonal or life functioning appeared to respond slower than problems relating to wellbeing or symptom distress. Kivlighan, Lin, Egan, Pickett and Goldberg (2018) however did not find a difference in domain rates when item dependency was controlled for on the BHM-20, but noted that early termination was associated with improvements on wellbeing but not on other domains (symptom distress or life functioning). Owen et al. (2016) describe that therapist effects explained some of the variance in rates of change in wellbeing and life functioning in their study, and Owen et al. (2015) noted that different sub-classes of people responded at different rates; notably the 'slow and steady' group had the slowest trajectories.

iv) Assessing the shape of change

Ten studies examined the shape of change. Five studies contrasted a DR model (aggregated) with a GEL model (stratified), and all found the GEL model a superior fit for the data, with Nielsen et al. (2016) noting that a combined model was better yet. One study described the shape of change based on visual inspection of plots of scores, and nine assessed the model fit of linear, log-linear, quadratic or cubic shapes of change.

Linear trends. A linear shape of change was found in seven studies under certain conditions. Barkham et al. (1996) noted that change tended to look linear when broken down into different symptoms, on individualised items, or when comparing sequences of RCSI rates. Reese, Toland and Hopkins (2011) used MLGM techniques and found that although a cubic term was significant, linear trends appeared to describe the data visually and more parsimoniously. Similarly, Erekson et al. (2015) found that linear terms provided the largest estimate.

Four studies comparing sub-groups found linear terms to be the best fit for those having longer treatment lengths. For example, Kivlighan et al. (2018) describe a linear pattern in clients having three or more sessions, as opposed to log-linear patterns evidenced in those having two or more. Falkenström et al. (2016) used MGLM and found a linear shape in a psychiatric sample who had longer treatment and slower rates of change, when compared with a quadratic trend seen in a primary care sample. Owen et al. (2016) also describe linear trends in those having longer treatments, whilst Owen et al. (2015) observed linear trends in the 'slow and steady' sub-group who had longer treatments (although note possible overlap in the latter two studies).

Log-linear trends. Four studies observed log-linear trends in the data. For example, Stulz et al. (2013) used LGCMs and found that log-linear terms outperformed linear in their sample, regardless of treatment length. As described above, however, Kivlighan et al. (2018) used MLMs and found that a log-linear shape was conditional on selection of those with at least two sessions rather than those with at least three. Owen et al. (2016) explored this further using MLMs, finding that a log-linear trend offered the best fit for the problem domains of wellbeing and symptom distress but not life functioning (which was quadratic), as well as for those having shorter treatments.

Nielsen et al. (2016) observed a log-linear trend in their data according to visual inspection and regression terms. They describe that a DR model with a NAC fits for shorter session lengths whilst the GEL fits longer treatment lengths and that a combined DR and GEL model is a better fit. However note that this represents an aggregate overview of the shape of change between samples rather than examining individual trajectories.

Cubic trends. A cubic trend was found to be the best fit in two studies. However Reese et al. (2011) stated that on inspection the trend was better described as

linear. Owen et al. (2015) also found an 'early-and-late' trend in their largest sub-class of clients, which would resemble a cubic trend.

Quadratic trends. Two studies found quadratic trends in some circumstances. As described above, Owen et al. (2016) found this trend on the problem domain of life functioning. Falkenström et al. (2016) found that a quadratic trend best described a primary care sample, whilst a linear term better described the psychiatric sample, who as discussed above tended to have higher risk and longer treatment lengths.

Meta-Analyses

Seven studies reported correlation coefficients for associations between improvement and total sessions (with one transformed from an odds ratio), and five reported correlation coefficients for associations between baseline scores and total sessions. Two meta-analyses were therefore carried out using correlation coefficients in random effects models, with sub-group analyses where appropriate.

Associations between improvement and total sessions. Five studies examined associations between either RCSI or RC and total sessions so their data were pooled for further examination. Two further studies used 'change in score' and one used regression on individual questionnaire domains rather, so they are not pooled but included for discussion.

i) ***All studies combined.*** Five studies were included with n=46,921 participants; all provided RCSI figures. A non-significant pooled effect size of r=-0.24 [95% CI -0.70, 0.36], p=0.2747 was found, suggesting no linear correlation between RCSI and total sessions. However this combined results from three studies showing negative or no correlations, with two studies that found small and large positive correlations, and high heterogeneity was

indicated Q(4)=18,655.94, p<.001), with I_2 of 100%. Publication bias analysis was non-significant according to the weight-function model likelihood ratio test $X_2(1)$=0.2178587, p= 0.64068, and the regression test for funnel plot asymmetry $t(3)$=1.0446, p=0.3730 (appendix C).

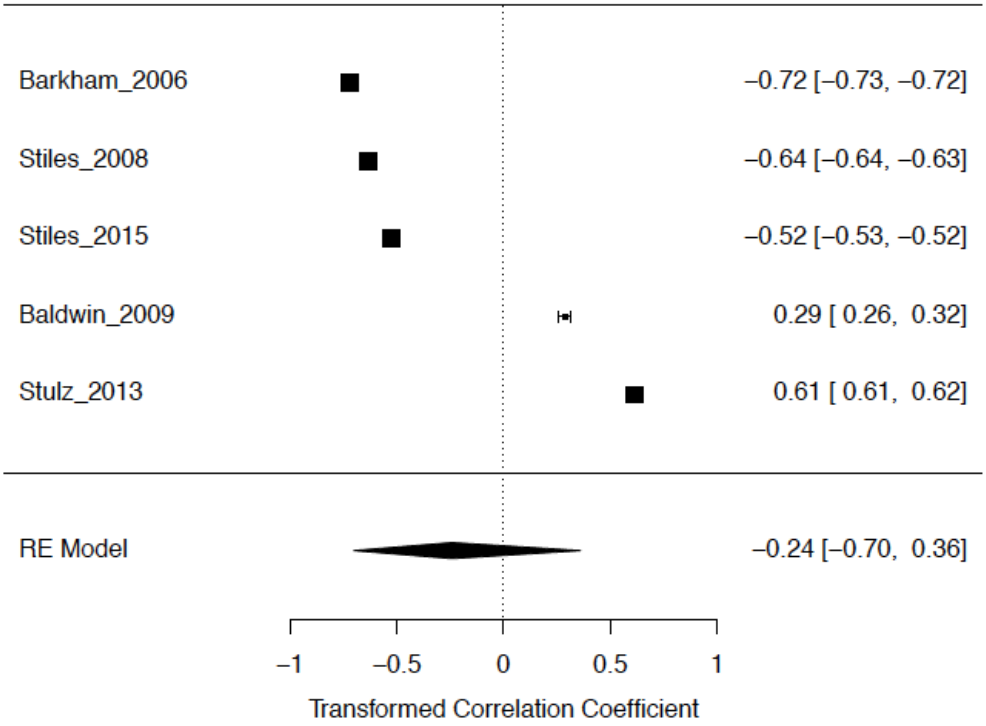


Figure 2. Forest plot for all studies using RCSI.

ii) ***Change in score and total sessions.*** Three further studies examined change scores and total sessions. Two found no association between change in score and total sessions using correlation analyses. One study reported regression figures broken down by domains, finding very small associations between change scores on different symptom domains and total sessions, with life functioning showing the least change (r_2=.014 on wellbeing, r_2=.021 on symptom distress, and r_2=.004 on life functioning).

Sub-group sensitivity analyses. It was of interest to examine whether different criteria influenced results. The figures for RCSI only are provided above, however further analyses were carried out examining the two studies reporting RC, as well as the effects of planned versus unspecified endings. Funnel plot asymmetry analyses were unreliable with only two studies and are not reported. These smaller sub-group comparisons are used to inform discussions on patterns in the data rather than to draw firm conclusions.

iii) ***RC only.*** Two studies were included with *n*=36,133 participants, and a pooled effect size of *r*=-0.14 [95% CI -0.57, 0.34], *p*=0.5574 was non-significant, suggesting no association between total sessions and reliable change. High heterogeneity was however indicated Q(1)=2024.07, *p*<.001), with I_2 of 100%. Publication bias analysis was non-significant according to the weight-function model $X_2(1)$=1.767006, *p*= 0.18375.

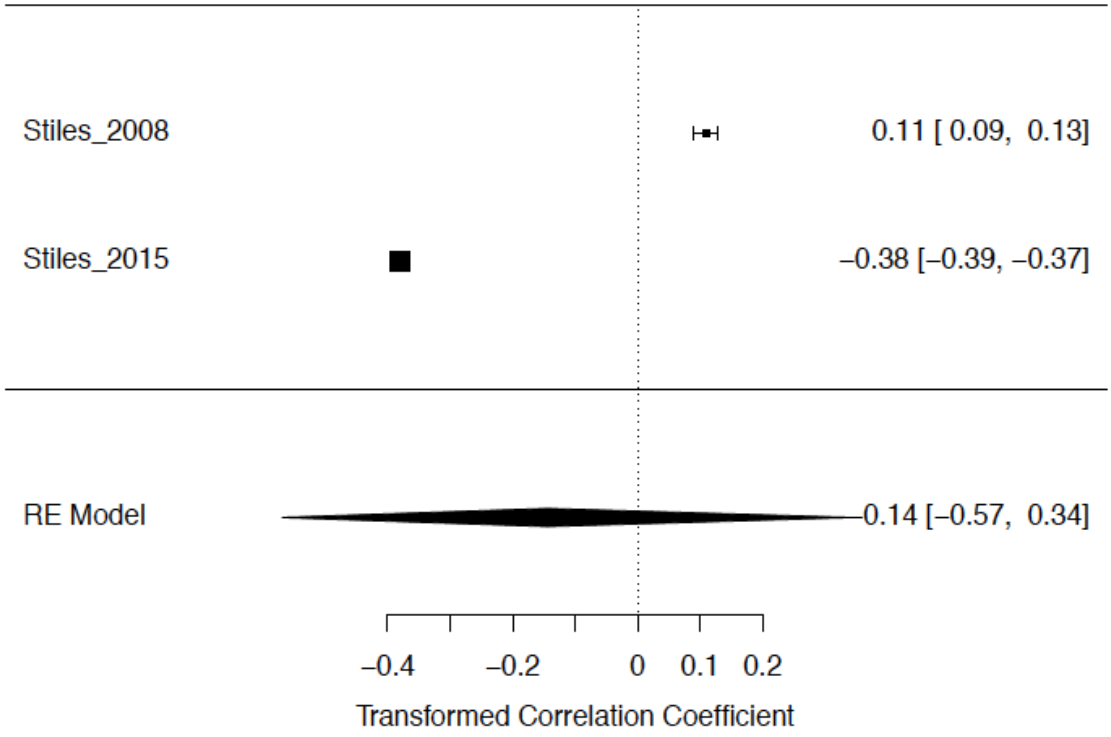


Figure 3. Forest plot for RC only

Further sub-group analyses were performed depending on whether the studies included planned endings exclusively or whether ending information was unspecified. Three studies included planned endings only, and two included endings unspecified (all reported RCSI, but two planned endings studies also reported RC). These were examined combined, as well as broken down by RCSI or RC criteria, to assess the possible influence of outcome definitions.

iv) ***Planned endings and RCSI*.** Three studies were included with n=37,605 participants. All three noted that some of the services included tended to limit therapy to six sessions (but not all), with flexibility to add more, and all reported RCSI figures. A significant large pooled effect size of r=-0.63 [95% CI -0.73, -0.51], p<.001 was found, suggesting a negative correlation between recovery and total sessions when planned endings only are included. However high heterogeneity was again indicated Q(2)=1546.61, p<.001), with I^2 of 99.9%. Although these studies all suggested a negative correlation between RCSI and total sessions, there were significant discrepancies between their effect sizes. Publication bias analysis was nonsignificant according to the weight-function model X^2(1)=4.570691e-07, p=0.99946 and funnel plot test t(1)= -2.3870, p=0.2526.

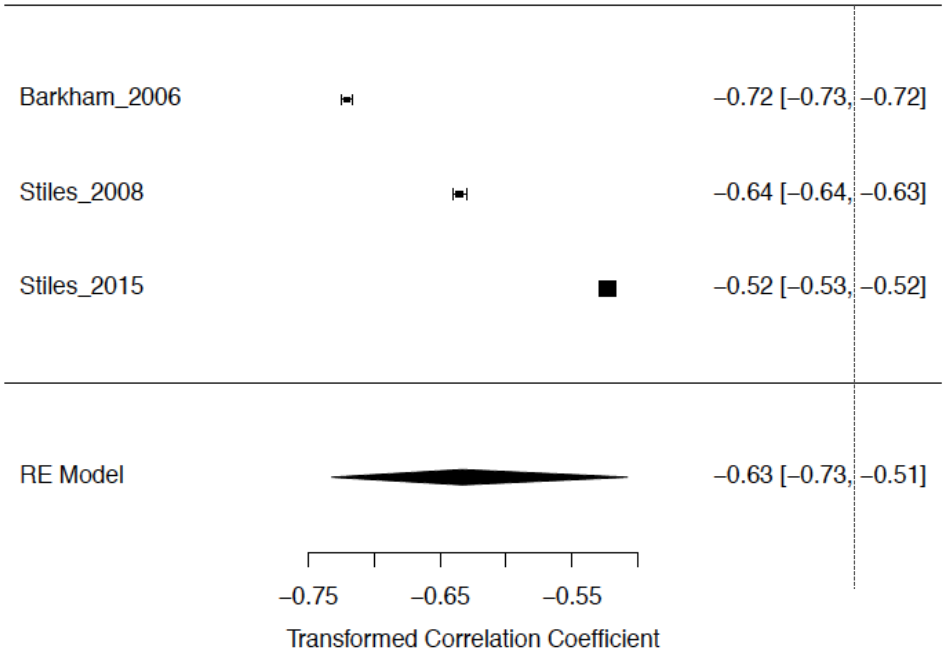


Figure 4. Forest plot for planned endings and RCSI

v) ***Planned endings and RC.*** Both studies included in the RC analysis above included planned endings so the figures are equivalent – no association was found between improvement and total sessions.

vi) ***Endings unspecified (all RCSI).*** Two studies were included n=9316. A significant moderate-large pooled effect size of r=0.47 [95% CI 0.10, 0.72], p=0.0419 was found. However high heterogeneity was indicated Q(1)=705.95 p<.001), with I^2 of 99.9%. Publication bias analysis was non-significant, with a weight-function test of $X^2(1)$=0.04923786, p=0.82439.

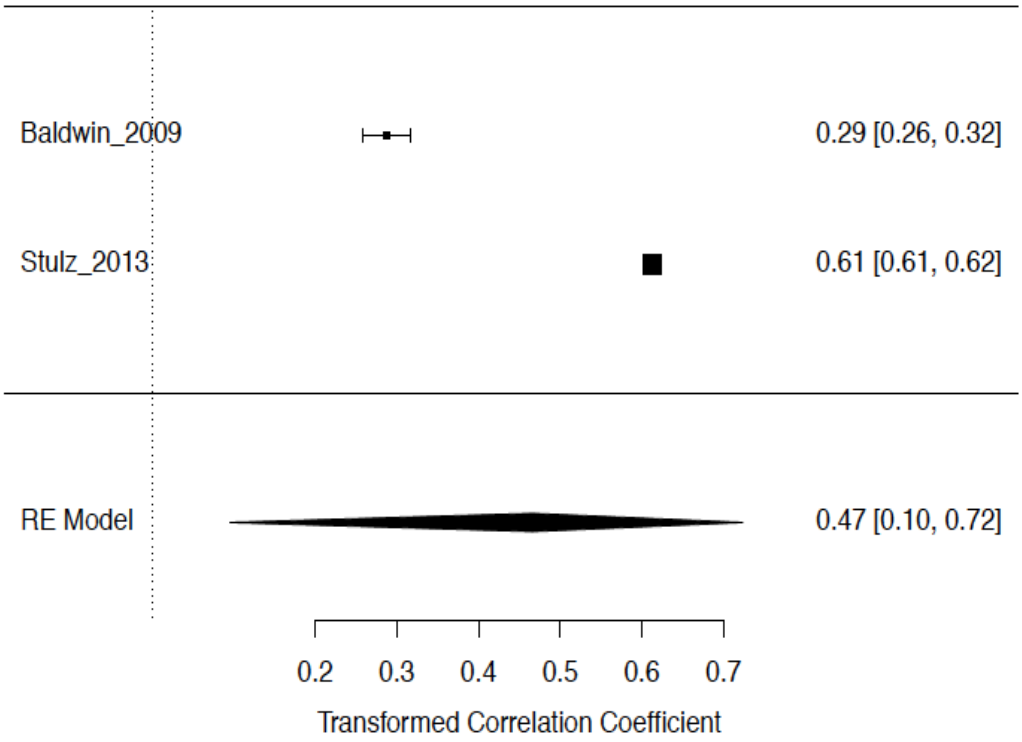


Figure 5. Forest plot for endings unspecified using RCSI

Figure 6 summarises these findings. When examining studies overall, no association was observed between improvement and total sessions, in line with the GEL model assumption. However sub-group analyses indicated that when the criteria of RCSI was used, planned endings produced a large negative correlation whereas unspecified endings produced a moderate-large positive correlation. When the criteria of change scores are used, no, or very small correlations are seen. High heterogeneity was observed in all cases apart from the change score analyses, however, and although sample numbers were high study numbers were low, making aggregate interpretations unreliable.

Of further note is that two studies used Spearman's rank order correlations, three studies used Pearson's linear correlations and two are unclear. It has been argued that the Pearson's method is less sensitive to non-linear associations, which may have affected results (Laerd Statistics, 2018). However also note that whilst Stulz et al.

(2013) found a strong positive association, Evans et al. (2017) did not, despite both using the Spearman's method. These results are also not taken in isolation but considered alongside alternative and non-linear approaches to examining the GEL.

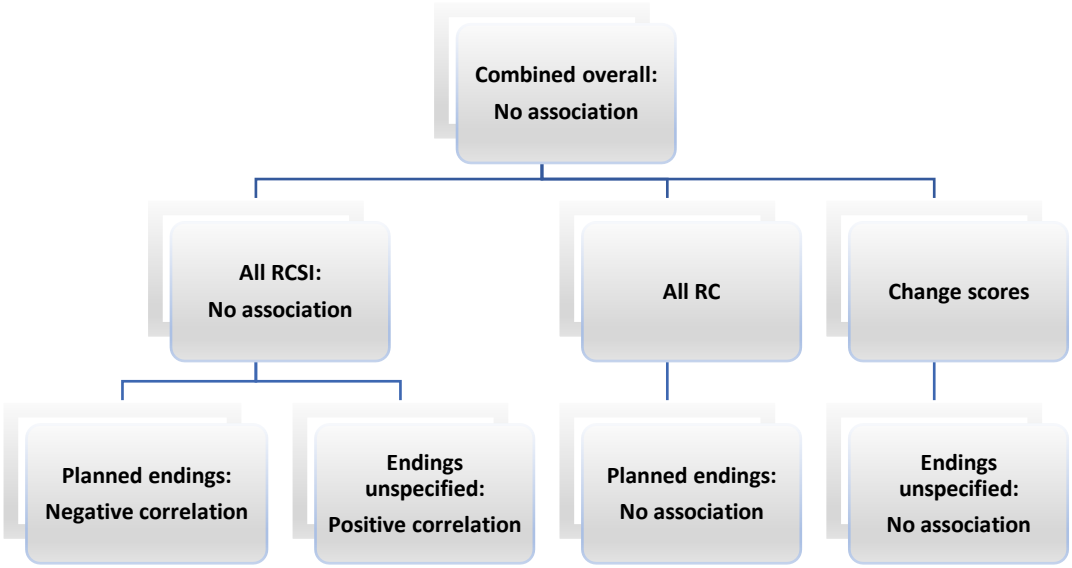


Figure 6. Sub-group comparisons of associations between improvement and total sessions.

Correlating initial symptom severity with total sessions. Five studies reported positive correlations between baseline symptom scores and total sessions attended. A significant small pooled effect size of r=0.15 [95% CI 0.08, 0.22], p<.001 was found, suggesting that higher baseline scores were associated with higher doses of total sessions. This provides support for the GEL argument that different people need different amounts of therapy, with individual need associated with increased sessions. However high heterogeneity was indicated Q(4)=83.20, p<.001), with I^2 of 95.2%. Publication bias analysis was non-significant according to the weight-function $X^2(1)$=1.078592, p= 0.29901, and funnel plot test t=1.4059 p=0.2544 (appendix C).

Two of these studies also found strong positive correlations when using mean rather than individual baseline scores (r=.93 and r=.58, p<.001).

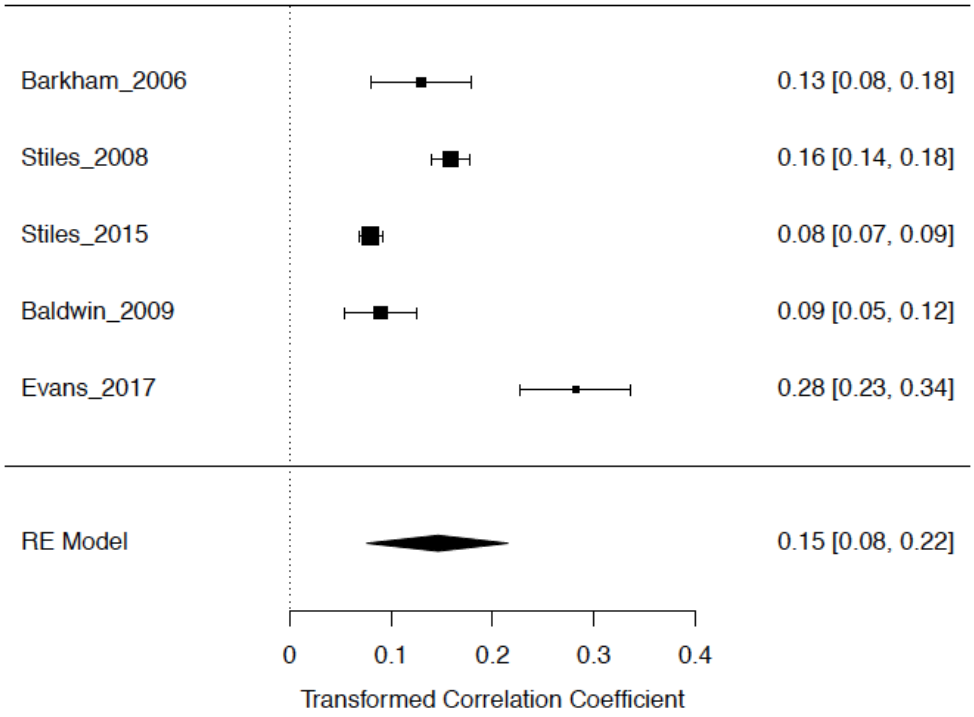


Figure 7. Forest plot for baselines correlated with total sessions.

Discussion

The research questions guiding this review were to understand whether different sub-groups of people respond to treatment at different rates in line with the GEL, and whether the shape of change is linear or non-linear. The studies approached this through examining whether improvement was associated with dose, whether baseline was associated with dose, whether rates of change differed, and through analysing the shape of change.

Main Findings

Associations between improvement and total sessions: completers versus non-completers. In support of the GEL model, no overall relationship was suggested

between improvement and total sessions, and this held whether the criteria of RCSI or RC was used. Different findings were observed depending on whether studies included or excluded unplanned endings, however. For example, when planned endings only were examined using the stricter criteria of RCSI, a large negative correlation was found, where those with higher session numbers were less likely to see RCSI. Conversely, a moderate-large positive correlation between RCSI and total sessions was found in studies that included unplanned endings. In other words, more sessions led to more recovery in these circumstances. This conflict was captured in Nielsen et al.'s study, where the only way that total sessions could predict change (DR) adequately was in a model that also allowed the possibility of change predicting sessions (GEL).

When unplanned endings are included, however, samples are likely to include those who terminate therapy early before criteria for improvement have been met. For example, Kivlighan et al. (2018) discussed that unplanned endings in their study occurred due to people making fast gains on wellbeing but not on other items. In practice, there may be a variety of reasons for why someone may terminate therapy early, including those who make poor progress (Delgadillo et al. 2014). In these studies, even though patients were stratified into sub-groups having the same treatment length, the average amount of improvement for those having fewer sessions appears lower because it includes those who drop-out before reaching improvement criteria.

In a sense, this is a further artefact of aggregation, where the effects of dose are not being examined from a within-subjects perspective but interpreted across different samples. Nonetheless, the majority of datasets will include non-completers so it could be that a combined DR and GEL is a good way to describe population aggregate patterns, whilst bearing in mind that this is not necessarily a reflection of individual responses to treatment.

Baseline symptom severity and increased total sessions. Seven out of eight studies that examined associations between baseline symptoms and total sessions found a small, positive correlation, where higher scores indicated longer treatment lengths. This supports the GEL proposal that different people may require different treatment lengths and links to previous research on patient profiling, which suggests that different classes of people may require different treatment intensities (Delgadillo, Moreea, & Lutz, 2016; Delgadillo, Huey, Bennett, & McMillan, 2017).

Rates of change are faster in those having fewer sessions. The concept of differential responses was further supported in studies examining the rates of change, with fewer total sessions being associated with steeper slopes or faster rates of change in all nine studies. On the one hand this is challenging for the DR perspective, because it suggests that there is not an average amount or dose that suits everyone. However this does not imply that unlimited treatment is beneficial. For example, DR findings suggest that the majority of people will have responded within defined treatment lengths dependent on the setting and nature of the problem (Robinson et al., 2019). Similarly, the above findings on negative correlations between treatment length and improvement suggest that some people may not see any further benefit from having increased sessions. Therefore it seems more likely that different people respond at different rates, within certain boundaries. The reasons for such differential response rates may be individual, but some of the studies in this review also highlighted issues such as therapist effects (Owen et al., 2016), session frequency (Erekson et al., 2015; Reese et al., 2011) and modality (Nielsen et al., 2016).

Mixed shapes of change. A variety of patterns were seen regarding the shape of change, however emerging patterns were noted. Of the nine studies examining shape, seven studies found linear trends, with four of these linking linear trends to longer

treatments when sub-groups were examined. Curvilinear responses were found in four studies, with one finding curvilinear responses regardless of treatment length, one finding the shape on sub-sets of symptoms, and two at shorter treatment lengths; although one of these observations were made on average group comparisons rather than within-subjects trends.

A cubic trend was only preferred overall in one study, however a trend similar to a cubic pattern was also observed in Owen et al.'s (2015) study in a sub-class of people where there was early change, a plateau then late change. This is interesting as it may add a caveat onto the previous discussion that not all people may benefit from longer treatments, where some may see a plateau and then make further progress. Quadratic trends were seen in two studies that broke analyses down into different sub-classes or problem domains, where the pattern may reflect progress on slower to respond problems such as life functioning (although Kivlighan et al. 2018 challenged the notion of differential response rates on domain items when dependency was controlled for).

In sum, these findings are in line with a growing body of literature that has identified different sub-classes of response (Delgadillo et al., 2014; Lutz et al., 2012; Rubel et al., 2015). Some emerging patterns were present in the data, suggesting that progress for more severe difficulties in longer treatments may look linear, whilst shorter treatments or responses on different symptoms may be log-linear or curvilinear. However within these broader patterns, there may also be sub-classes of people who follow different trajectories, which may reflect responses to different problems, different phases of therapy, or other individual or external influences on change.

Thus, whilst there may be some instances in which the shape of change follows a decelerating trend, this effect may have been inflated in the literature by the aggregation across groups, and the shape cannot be considered to reflect average

individual curvilinear responses. Even though modelling techniques were used to examine individual trajectories, these still represent an average of lots of different individual growth curves and again may potentially mask variability. For example, initial curves are strongly influenced by fast responders or early drop-out, whilst later curves or tails are strongly influenced by slower responders.

Limitations

Most of the review studies were subject to limitations found in naturalistic databases, that are either inherent or commonly found. Overlap therefore exists between the individual study limitations and the review's limitations. Study limitations that have not already been described are discussed here first before review specific issues are identified below.

Missing data. The issue of missing data is problematic for examinations of responses to psychotherapy, as they are inherently made on those with completed outcome measures, even if they terminated early. Although missing data is often treated as missing at random for statistical purposes, the reasons for its absence are often not clear. Erekson et al. (2015) found that missing session data in their study were correlated with session frequency, total sessions and baseline symptom severity. Evans et al. (2017) showed that those with completed measures in their sample were more likely to be older, White British, and with lower baselines than those without. In other words there may be important information contained within missing data however this is not represented by studies dependent on completed outcomes data.

Gottfredson et al. (2014) illustrated that when SPMM methods were used to handle missing data, findings suggested that participants with faster recovery rates terminated therapy earlier, meaning that rates of change are generally underestimated according to traditional 'missing at random' models. They explain that over time small

bias may become substantial, which is a problem if we are making average interpretations about responses to treatment and want to understand or predict future patients' trajectories.

Missing client demographics, presenting problem and intervention characteristics. All studies bar one were retrospective database analyses and as such were reliant on the recording of demographic and treatment information by the included services. Although missing characteristics do not preclude the examination of responses to treatment, they may limit interpretations and considerations of how findings translate to other services and settings. Missing information may also limit possibilities for considering patterns in the data and forming new research hypotheses.

For example, it would be of particular interest for examining the GEL to know reasons for treatment ending, as well as whether interventions took a phased approach, working on different problems at different times. Some services may use of the concept of spaced learning, staggering penultimate sessions and using them for relapse prevention planning or consolidation. Greater change would not be expected during these phases, however this would contribute to an overall decelerating pattern of change. Other services however may be more limited and endings more abrupt. Factors such as these could help explain some of the differences between study findings on shapes of change. Indeed 'big data' can be extremely beneficial for facilitating advances in our understanding of patterns of psychotherapy response; however it can also mask heterogeneity across settings unless characteristics are well documented.

Review limitations. Factors such as these may explain the high heterogeneity found between studies, even when they were isolated by factors such as outcome and inclusion criteria. In addition, although sample numbers were large, study numbers were small, meaning that the review was unable to draw strong conclusions about any effects

found in the meta-analyses. It was necessary to limit searches to the English language, however this can also introduce bias to reviews where only papers selected for submission to English language journals are included. Similarly, the grey literature was excluded for pragmatic reasons because the topic of the review is specific, large numbers of studies were not expected, and including grey literature could lead to the inclusion of studies that have not been reviewed or assessed for methodological quality. This was considered problematic where methodological expertise is required to evaluate this field. Given the high response rate from authors in the field, it is unlikely that substantial literature has been missed on the topic, but it is nonetheless possible.

Strengths

This is the first review to synthesise literature on the GEL, including the first time that GEL model assumptions have been tested using meta-analytic methods. This is important given the potential importance of this broader field in influencing the development of services and wellbeing of clients. Comprehensive and systematic searches were performed, and bias ratings were carried out by two reviewers with discussion on the items. These discussions of bias were particularly helpful as they provided a platform for understanding the patterns across the different studies. Although there was high heterogeneity and aggregate conclusions are not considered appropriate, the review has been able to identify patterns emerging in the evidence base and offer possible explanations for how both GEL and DR patterns have been observed to date.

Such heterogeneity can be viewed as undesirable, making it hard to reach firm quantitative conclusions. However if all studies had taken a similar methodological approach with rigidly defined criteria, this in turn would have created a narrow and biased review. The variety of approaches taken by the studies in this review have provided possibilities to learn about the GEL from different angles. Further, a

discussion of the limitations that most naturalistic studies face has also highlighted some useful areas for future theoretical and research development.

Theoretical Implications and Future Research

Several key theoretical questions have emerged from this review. For example, if some people respond more rapidly to therapy than others, it is of interest to know if we can identify their profiles. If so, it could be possible to prescribe low intensity and low-cost treatments to them, making better use of resources. Similarly, if other people are gradual responders who need more intensive treatment, it would be important to identify them early and match them to more intensive treatment. Research has made some progress in this area (see Delgadillo et al. 2016; DeRubeis et al., 2014); however more research is needed to understand how best to stratify care in the future. Some of the papers in this review have also highlighted other influences on rates of change, such as session frequency, therapy modality, and therapist effects (see also Goldberg et al., 2018). Future research in these areas within the context of providing stratified care would be beneficial.

Given the limitations discussed in the review, there are several relatively simple changes that services and/or researchers could make to enable better research in the field, such as including coding for treatment endings where possible. This seems to be a critical part of assessing DR and GEL models and the shape of change. It would also be helpful to have more information on service constraints or culture. For example, do services on the whole tend to end treatment abruptly or is there flexibility for clinicians and clients to responsively regulate, and are maintenance or consolidation sessions the norm. Slightly more complicated but potentially useful is to know whether therapy can be defined in terms of phases (Howard, Lueger, Maling, & Martinovich, 1993). Further research on shapes of change given the above might also be helpful for clinicians

attempting to interpret progress. For example, it would be important to understand whether someone has plateaued and is unlikely to make further progress, or whether they are an early and late responder.

Nine of the review studies used data from university counselling centres in the US, and in the UK the majority of the research came from primary care. It would therefore be of interest to understand how these findings hold in other potentially more complex samples, given the emerging trends for differences depending on baseline profile. Finally it would be of interest to understand further what a GEL means to people. For example, Kivlighan et al. (2018) noted that some people made progress on aspects such as wellbeing and terminated there, before making progress on other symptoms. A question for future research is whether this constitutes a GEL and if so how can this be captured in research findings. This could make a case for the further bridging of nomothetic and idiographic outcomes in the field.

Clinical Implications

These findings suggest that services could be planned in a way where they can "responsively regulate" treatment in an empirically informed way, but within certain boundaries as supported by the dose-response literature. The studies in this review suggest that some patients might respond rapidly, and these may tend to have milder baseline severity, require shorter treatment lengths, and where change may be initially fast before tapering off as treatment goes on. However some respond more slowly, and these may be more likely to have higher baseline symptom severity, require longer treatment and see a more linear overall pattern of progress. Within this, there may also be sub-classes of people who see plateaus in the middle of therapy before making gains again, and this may reflect their response to particular problems, such as life functioning. However more research is needed into shapes of change and this remains

hard to predict at outset. Clinical conversations between client-therapist and therapist-supervisor should therefore explore the needs of clients and any unexpected patterns of response. Services could also consider the effects of therapy modality and session frequency when considering stratified care.

Overall this review suggests that different people require different doses, according to their needs and problems. However, there's no evidence that indefinite or extremely lengthy treatments are necessary. There is an upper boundary where most responders are identified, within primary care and university counselling settings at least (Robinson et al., 2019). Insights from the GEL and DR models could be integrated into a coherent concept of "boundaried responsive regulation". Different people require different doses of treatment, but this responsive regulation of treatment duration occurs within predictable dose boundaries. Services could therefore offer flexible appointments for patients with different needs, within the confines of the dose-response boundaries for that clinical population. An even better approach in future could be stratified care: offer low-cost treatments for rapid responders, and high-cost treatments for gradual responders.

Conclusions

In sum, this review has illustrated that the GEL and DR perspectives are not entirely incompatible and both offer valid insights. The GEL model better represents the heterogeneity in responsiveness across different patients. However, from a population perspective, even if people responsively regulate and exhibit different change trajectories, most responders are identified within a finite number of sessions within these contexts; eventually there may be people remaining in lengthy treatments who do not make improvements according to typical outcome criteria. Both models can co-exist and inform treatment planning within a concept of “boundaried responsive regulation”.

References

*Indicates review paper

Altman, D. G. (1999). *Practical statistics for medical research*. New York, NY: Chapman & Hall/CRC Press.

*Baldwin, S. A., Berkeljon, A., Atkins, D. C., Olsen, J. A., & Nielsen, S. L. (2009). Rates of change in naturalistic psychotherapy: Contrasting dose-effect and good-enough level models of change. *Journal of Consulting and Clinical Psychology, 77*, 203–211. https://doi.org/10.1037/a0015235

*Barkham, M., Connell, J., Stiles, W., Miles, J., Margison, F., Evans, C., . . . La Greca, A. M. (2006). Dose-effect relations and responsive regulation of treatment duration: The good enough level. *Journal of Consulting and Clinical Psychology, 74,* 160-167. DOI:10.1037/0022-006X.74.1.160

*Barkham, M., Rees, A., Stiles, W., Shapiro, D., Hardy, G., & Reynolds, S. (1996). Dose-effect relations in time-limited psychotherapy for depression. *Journal of Consulting and Clinical Psychology, 64,* 927-935. http://dx.doi.org/10.1037/0022-006X.64.5.927

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. (2009). *Introduction to meta-analysis.* West Sussex, UK: John Wiley & Sons.

Castonguay, L., Barkham, M., Lutz, W., & McAleavey, A. (2013). Practice-oriented research: Approaches and applications. In M. Lambert (Ed.). *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 85-133). New Jersey, NJ: John Wiley & Sons, Inc.

Cherry G., & Dickson, R. (2014). Defining my review question. In A. Boland, G. Cherry, & R. Dickson (Eds.). *Doing a systematic review* (pp.17-33). London, UK: SAGE.

Critical Appraisal Skills Programme (2018). *Cohort study checklist.* Retrieved from: https://casp-uk.net/casp-tools-checklists/

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development : Official Journal of the Cognitive Development Society, 11,* 121–136. https://doi.org/10.1080/15248371003699969

Delgadillo, J., de Jong, K., Lucock, M., Lutz, W., Rubel, J., Gilbody, S…. & Mcmillan, D. (2018). Feedback-informed treatment versus usual psychological treatment for depression and anxiety: a multisite, open-label, cluster randomised controlled trial. *The Lancet Psychiatry, 5,* 564–72. https://doi.org/10.1016/S2215-0366(18)30162-7

Delgadillo, J., Huey, D., Bennett, H., & McMillan, D. (2017). Case complexity as a guide for psychological treatment selection. *Journal of Consulting and Clinical Psychology, 85,* 835-853. http://dx.doi.org/10.1037/ccp0000231

Delgadillo, J., McMillan, D., Lucock, M., Leach, C., Ali, S., & Gilbody, S. (2014). Early changes, attrition, and dose-response in low intensity psychological interventions. *British Journal of Clinical Psychology, 53*, 114–130. https://doi.org/10.1111/bjc.12031

Delgadillo, J., Moreea, O., & Lutz, W. (2016). Different people respond differently to therapy: A demonstration using patient profiling and risk stratification. *Behaviour Research and Therapy, 79,* 15-22. http://dx.doi.org/10.1016/j.brat.2016.02.003

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014). The personalized advantage index: translating research on prediction into individualized treatment recommendations. A demonstration. *PLoS ONE, 9,* e83875. http://dx.doi.org/10.1371/journal.pone.0083875.

Egger, M., Smith, G. D., Schneider, M., & Minder C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315,* 629. https://doi.org/10.1136/bmj.315.7109.629

*Erekson, D. M., Lambert, M. J., & Eggett, D. L. (2015). The relationship between session frequency and psychotherapy outcome in a naturalistic setting. *Journal of Consulting and Clinical Psychology*, *83*, 1097–1107. https://doi.org/10.1037/a0039774

*Evans, L. J., Beck, A., & Burdett, M. (2017). The effect of length, duration, and intensity of psychological therapy on CORE global distress scores. *Psychology and Psychotherapy: Theory, Research and Practice, 90,* 389-400. https://doi.org/10.1111/papt.12120

*Falkenström, F., Josefsson, A., Berggren, T., & Holmqvist, R. (2016). How much therapy is enough? Comparing dose-effect and good-enough models in two different settings. *Psychotherapy*, *53*, 130–139. https://doi.org/10.1037/pst0000039

Field, A. (2018). Discovering statistics using IBM SPSS statistics (5th ed.). London, UK: SAGE Publications Ltd.

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical & Statistical Psychology, 63*, 665-694. https://doi.org/10.1348/000711010X502733

Goldberg, S. B., Hoyt, W. T., Nissen-Lie, H. A., Nielsen, S. L., & Wampold, B. E. (2018). Unpacking the therapist effect: Impact of treatment length differs for high- and low-performing therapists. *Psychotherapy Research, 28*, 532-544. https://doi-org.sheffield.idm.oclc.org/10.1080/10503307.2016.1216625

*Gottfredson, N. C., Bauer, D. J., Baldwin, S. A., & Okiishi, J. C. (2014). Using a shared parameter mixture model to estimate change during treatment when termination is related to recovery speed. *Journal of Consulting and Clinical Psychology, 82*, 813-827. http://dx.doi.org/10.1037/a0034831

Hamilton, W. (2017). Package 'MAVIS' (Version 1.1.3). Retrieved from: http://kylehamilton.net/shiny/MAVIS/

Higgins, J. P. T., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions* (version 5.1.0). The Cochrane Collaboration. Retrieved from: www.handbook.cochrane.org.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*, 1539-1558. https://doi.org/10.1002/sim.1186

Howard, K.I., Kopta, S.M., Krause, M.S., & Orlinksy, D.E. (1986). The dose-effect relationship in psychotherapy. *American Psychologist, 41*, 159-164. http://dx.doi.org/10.1037/0003-066X.41.2.159

Howard, K. I., Lueger, R. J., Maling, M. S., & Martinovich, Z. (1993). A phase model of psychotherapy outcome: Causal mediation of change. *Journal of Consulting*

and Clinical Psychology, 61, 678-685. http://dx.doi.org/10.1037/0022-006X.61.4.678

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12-19. http://dx.doi.org/10.1037/0022-006X.59.1.12

Kivlighan, D. M., Lin, Y. J., Egan, K. P., Pickett, T., & Goldberg, S. B. (2018). A further investigation of the good-enough level model across outcome domains and termination status. *Psychotherapy, (advance online publication)*. http://dx.doi.org/10.1037/pst0000197

Laerd Statistics (2018). *Pearson's product-moment correlation using SPSS Statistics. Statistical tutorials and software guides.* Retrieved from: https://statistics.laerd.com

Lambert, M., Shimokawa, K., & Hilsenroth, Mark J. (2011). Collecting Client Feedback. *Psychotherapy, 48*, 72-79. http://dx.doi.org/10.1037/a0022238

Lutz, W., Ehrlich, T., Rubel, J., Hallwachs, N., Röttger, M., Jorasz, C., . . . Tschitsaz-Stucki, A. (2012). The ups and downs of psychotherapy: Sudden gains and sudden losses identified with session reports. *Psychotherapy Research, 23,* 1-11. https://doi-org.sheffield.idm.oclc.org/10.1080/10503307.2012.693837

Lutz, W., Martinovich, Z., & Howard, K. I. (1999). Patient profiling: An application of random coefficient regression models to depicting the response of a patient to outpatient psychotherapy. *Journal of Consulting and Clinical Psychology, 67*, 571-577. http://dx.doi.org/10.1037/0022-006X.67.4.571

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., The PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6,* e1000097. https://doi.org/10.1371/journal.pmed.1000097

*Nielsen, S. L., Bailey, R. J., Nielsen, D. L., & Pedersen, T. R. (2016). Dose response and the shape of change. In S. Maltzman, *The Oxford handbook of treatment processes and outcomes in psychology: A multidisciplinary, biopsychosocial approach* pp.465-496. Oxford, UK: Oxford University Press.

*Owen, J., Adelson, J., Budge, S., Wampold, B., Kopta, M., Minami, T., & Miller, S. (2015). Trajectories of change in psychotherapy. *Journal of Clinical Psychology, 71,* 817-827. https://doi.org/10.1002/jclp.22191

*Owen, J. J., Adelson, J., Budge, S., Kopta, S. M., & Reese, R. J. (2016). Good-enough level and dose-effect models: Variation among outcomes and therapists. *Psychotherapy Research*, *26*, 22–30. https://doi.org/10.1080/10503307.2014.966346

*Reese, R. J., Toland, M. D., & Hopkins, N. B. (2011). Replicating and extending the good- enough level model of change: Considering session frequency. *Psychotherapy Research, 21,* 608–619. https://doi.org/10.1080/10503307.2011.598580

Robinson, L., Delgadillo, J. & Kellett, S. (2019). The dose-response effect in routinely delivered psychological therapies: A systematic review. *Psychotherapy Research. https://doi.org/10.1080/10503307.2019.1566676*

Rubel, J., Lutz, W., & Schulte, D. (2015). Patterns of change in different phases of outpatient psychotherapy: A stage-sequential pattern analysis of change in session reports. *Clinical Psychology and Psychotherapy, 22,* 1-14. http://dx.doi.org/10.1002/cpp.1868

*Stiles, W. B., Barkham, M., Connell, J., & Mellor-Clark, J. (2008). Responsive Rrgulation of treatment duration in routine practice in United Kingdom primary care settings: Replication in a larger sample. *Journal of Consulting and Clinical Psychology, 76,* 298-305. doi:10.1037/0022-006X.76.2.298

*Stiles, W. B., Barkham, M., & Wheeler, S. (2015). Duration of psychological therapy: relation to recovery and improvement rates in UK routine practice. *British Journal of Psychiatry, 207*, 115-122. https://dx.doi.org/10.1192/bjp.bp.114.145565

Stiles, W. B., Honos-Webb, L., & Surko, M. (1998). Responsiveness in psychotherapy. *Clinical Psychology: Science and Practice, 5,* 439-458. https://doi.org/10.1111/j.1468-2850.1998.tb00166.x

*Stulz, N., Lutz, W., Kopta, S. M., Minami, T., & Saunders, S. M. (2013). Dose-effect relationship in routine outpatient psychotherapy: Does treatment duration matter? *Journal of Counseling Psychology, 60*, 593–600. https://doi.org/10.1037/a0033589

Vevea, J.L., & Hedges, L.V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika, 60,* 419-435. http://dx.doi.org/10.1007/BF02294384

Appendix A

Example search strategy and results for PsycINFO, Medline and Scopus.

PsycINFO OvidSP 1806 – April week 2 2019, date searched 13.04.19

Notes: all limited by English language and human research.

.mp. = [mp=title, abstract, heading word, table of contents, key concepts, original title, tests & measures]

1. (good enough level) or (good-enough level).mp.
2. (dose effect or dose-effect or dose response or dose-response).mp.
3. "responsive regulation".mp.
4. "rate of change".mp.
5. Treatment duration.mp. or exp treatment duration/
6. Exp treatment outcomes/ or treatment outcome*.mp.
7. Exp group psychotherapy/ or exp psychotherapy/ or psychotherap*.mp.
8. Cognitive behavio?r therapy.mp. or exp behavior therapy/ or exp cognitive behavior therapy/ or exp cognitive therapy/
9. Exp counselling/ or psychological therapy.mp.
10. 1 or 2 or 3 or 4 or 5
11. 7 or 8 or 9
12. 6 and 10 and 11
13. Limit 12 to human and English language

N=348

OvidSP Medline [R] and Epub ahead of print, In-process & other non-indexed citations, daily and versions – 1946 to April Week 1 2019.

1. (good enough level) or (good-enough level).mp.
2. (dose effect or dose-effect or dose response or dose-response).mp.
3. "responsive regulation".mp.
4. "rate of change".mp.
5. "Treatment length" or "treatment duration" or "therap* dose" or *therap* length" or "therap* duration".mp.
6. Exp treatment outcome/ or treatment outcome*.mp.
7. Exp Psychotherapy, Group/ or exp Psychotherapy/ or psychotherap*.mp.
8. Cognitive behavio?r therapy.mp. or exp cognitive behavioral therapy/
9. "psychological therap* or "counsel?ing"
10. 1 or 2 or 3 or 4 or 5
11. 7 or 8 or 9
12. 6 and 10 and 11
13. Limit 12 to human and English language

N=456

Ovid did not find subject headings on treatment duration or psychological therapy so key terms scoped.

Note: [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonym

Scopus: no date limit – oldest paper scoped 1965 – to 14.04.19.

(TITLE-ABS-KEY ("Good enough level" OR "Good-enough level") OR TITLE-ABS-KEY ("dose effect" OR "dose-effect" OR "dose response" OR "dose-response") OR TITLE-ABS-KEY ("responsive regulation") OR TITLE-ABS-KEY ("rate of change") OR TITLE-ABS-KEY ("treatment duration" OR "treatment length" OR "treatment dose" OR "therap* length" OR "therap* dose" OR "therap* duration") AND TITLE-ABS-KEY (outcome* OR response OR change OR improv*) AND TITLE-ABS-KEY ("psychotherap*" OR "psychological therap*" OR "counselling or counseling OR "cognitive behavio* therap*" OR "CBT")) AND (LIMIT-TO (SUBJAREA , "PSYC") OR LIMIT-TO (SUBJAREA , "SOCI") OR LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "MULT")) AND (LIMIT-TO (LANGUAGE , "English"))

N=1495

Appendix B

Data Extraction and Risk of Bias Assessments based on CASP Cohort Study Checklist (2018) and Cochrane Guidance (2011).

Data Extraction Questions

1	Title
2	Aims
3	Setting
4	Sample N
5	Demographics
6	Inclusion criteria
7	Exclusion criteria
8	Presenting problem
9	Intervention
10	Outcome measures
11	Outcome criteria
12	Design
13	Method
14	Analysis of GEL
15	Treatment duration
16	Findings
17	Effect sizes
18	Conclusions
19	Include?

Risk of Bias Questions

1	Are the study aims clear?
2	Was recruitment acceptable?
3	Was the intervention accurately measured?
4	Was the outcome accurately measured?
5	Did the authors ID confounds?
6	Did they account for confounds in their design/analysis?
7	Are the results clearly reported?
8	How precise are the results?
9	Are the results credible?
10	Are the implications for practice credible?
11	Are results selectively reported?
12	Is there incomplete or unaccounted for outcome data?
13	Any other bias?

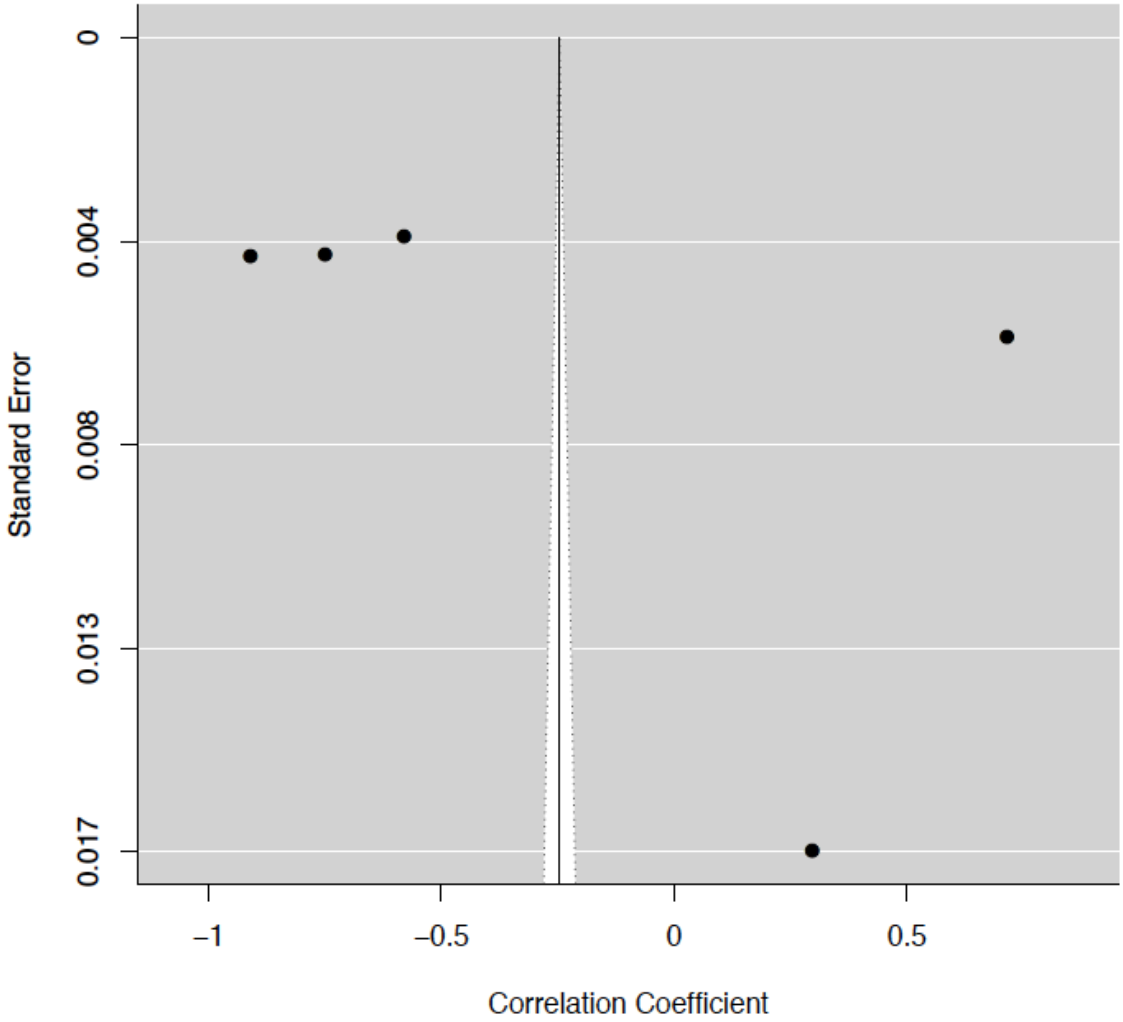
Risk of Bias Assessment

		Q1		Q2		Q3		Q4		Q5		Q6		Q7		Q8		Q9		Q10		Q11		Q12		Q13		Q14	
		Are the study aims clear?		*Was recuitment acceptable?*		*Was the intervention accurately measured?*		*Was the outcome accurately measured?*		*Do the authors ID all confounds?*		*Do they account for confounds in their design/analysis?*		*Are the results clearly reported?*		*How precise are the results?*		*Are the results credible?*		*Are the implications for practice credible?*		*Are results selectively reported?*		*Is there incomplete or unaccounted for outcome data?*		*Other bias?*		*Overall bias: Low, high, unclear (overall rating = median)*	
1	Baldwin et al. 2009	✓	0	✓	0	✓	0	✓	0	✓	0	!	1	✓	0	✓	0	✓	0	✓	0	✓	0	!	1	✓	0	✓	11
2	Barkham et al. (1996)	✓	0	✓	0	✓	0	✓	0	!	1	!	1	✓	0	!	1	✓	0	✓	0	✓	0	✓	0	✓	0	✓	10
3	Barkham et al. (2006)	✓	0	✓	0	✓	0	✓	0	✓	0	!	1	✓	0	!	1	✓	0	✓	0	✓	0	!	1	✓	0	✓	10
4	Erekson et al. (2015)	✓	0	✓	0	✓	0	✓	0	!	1	!	1	✓	0	✓	0	✓	0	✓	0	✓	0	!	1	✓	0	✓	10
5	Evans et al. (2017)	✓	0	✓	0	✓	0	✓	0	✓	0	✓	0	✓	0	!	1	✓	0	✓	0	✓	0	!	1	✓	0	✓	11
6	Falkenstrom et al. (2016)	✓	0	✓	0	✓	0	✓	0	!	1	!	1	✓	0	✓	0	✓	0	✓	0	✓	0	✗	2	✓	0	✓	10
7	Gottfredson et al. (2014)	✓	0	✓	0	✓	0	✓	0	✓	0	✓	0	✓	0	✓	0	✓	0	✓	0	✓	0	✓	0	✓	0	✓	13
8	Kivlighan et al. (2018)	✓	0	✓	0	✓	0	✓	0	✓	0	!	1	✓	0	✓	0	✓	0	✓	0	✓	0	✗	2	✓	0	✓	11
9	Nielsen et al. (2016)	✓	0	✓	0	✓	0	✓	0	!	1	!	1	✓	0	✓	0	✓	0	✓	0	✓	0	!	1	✓	0	✓	10
10	Owen et al. (2015)	✓	0	✓	0	✓	0	✓	0	!	1	!	1	✓	0	✓	0	✓	0	✓	0	✓	0	✗	2	✓	0	✓	10
11	Owen et al. (2016)	✓	0	✓	0	✓	0	✓	0	!	1	!	1	✓	0	✓	0	✓	0	✓	0	✓	0	✗	2	✓	0	✓	10
12	Reese et al. (2011)	✓	0	✓	0	✓	0	✓	0	✓	0	✓	0	✓	0	✓	0	✓	0	✓	0	✓	0	!	1	✓	0	✓	12
13	Stiles et al. (2008)	✓	0	✓	0	✓	0	✓	0	✓	0	✓	0	✓	0	!	1	✓	0	✓	0	✓	0	!	1	✓	0	✓	11
14	Stiles et al. (2015)	✓	0	✓	0	✓	0	✓	0	✓	0	✓	0	✓	0	!	1	✓	0	✓	0	✓	0	!	1	✓	0	✓	11
15	Stulz et al. (2013)	✓	0	✓	0	✓	0	✓	0	!	1	!	1	✓	0	✓	0	✓	0	✓	0	✓	0	!	1	✓	0	✓	10

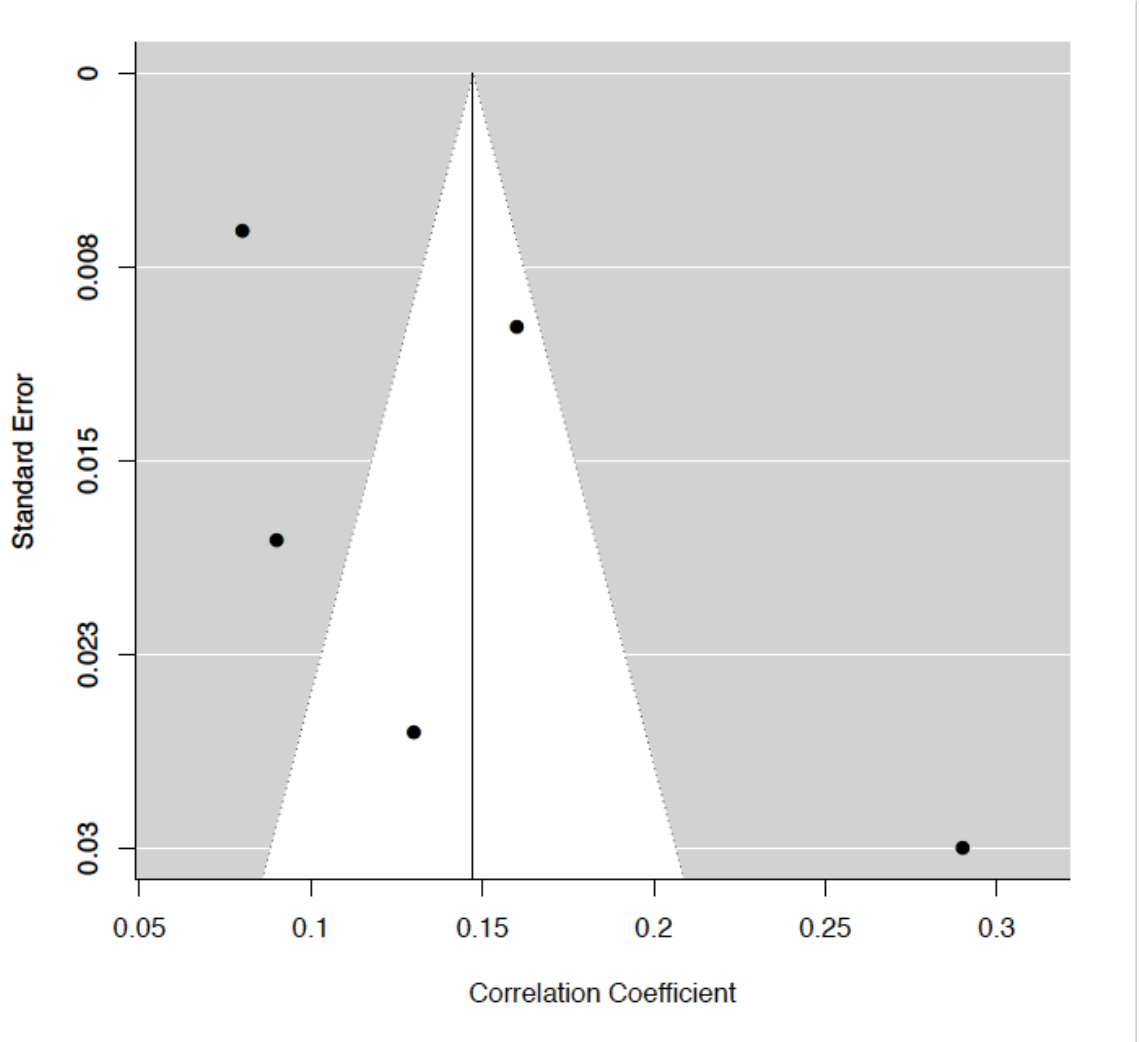
Appendix C

Funnel Plots for Primary Meta-Analyses

Funnel Plot for Associations Between Improvement and Total Sessions



Funnel Plot for Associations Between Baselines and Total Sessions



Section Two: Empirical Study

Title: THE DEVELOPMENT OF A DYNAMIC PROGRESS FEEDBACK SYSTEM TO GUIDE PSYCHOLOGICAL TREATMENT IN PRIMARY CARE

Short title: DEVELOPMENT OF A DYNAMIC PROGRESS FEEDBACK SYSTEM

Claire E. Bone

University of Sheffield

This page is left intentionally blank

Abstract

Objectives

This study aimed to develop dynamic progress feedback models, combining initial profile information from the Leeds Risk Index ([LRI] Delgadillo, Moreea, & Lutz, 2016) with weekly progress scores to provide personalised prognoses of recovery. Sub-aims were to assess generalisability, and to examine whether complex models outperformed parsimonious ones.

Design

A retrospective database analysis was used to construct the predictive models in one Improving Access to Psychological Therapies (IAPT) dataset, followed by cross-validation in a new IAPT dataset.

Methods

Models of increasing complexity were constructed, using backward elimination to retain significant predictors. Five predictors were used: baseline score, current session score, cumulative risk sums, cumulative individual standard deviations, and LRI profile group. Models were compared on how much variance they explained, and AUC, Kappa and Brier scores used for cross-validation.

Results

The models showed good predictive ability and cross-validated well in a new sample (e.g. explaining 39% of variance with an AUC of .775 by session four at low intensity). At high intensity treatment, the most complex model was superior at sessions one-to-six. At cross-validation, the more complex models saw higher scores, however differences between models were not statistically significant.

Conclusions

It was possible to build dynamic prediction systems that cross-validated in a new sample. Although complex models performed slightly better, this was not statistically significant at cross-validation. The question of whether to incorporate more complexity would therefore be a service-led decision. Further cross-validation and development into a clinical tool is intended.

Keywords: “outcome feedback”, “psychotherapy outcomes”, “patient profiling”, “predictive modelling”, “outcomes prediction”

Practitioner Points

- Dynamic prediction systems can be used as a form of “SatNav” system to provide personalised prognoses of recovery, which update as new weekly progress scores are added
- These systems could be used to support clinician judgement about which cases to take to supervision and to prompt clinical conversations and action to ameliorate risk
- More complex models including the LRI may be helpful at higher intensity therapies, whilst parsimonious models are sufficient at low intensity. Preference may depend on service capability for administering the LRI

Key Limitations

- The models cannot capture all influences on outcomes
- Although models cross-validated well in a new sample, confidence intervals overlapped so it is unclear how much benefit arises from extra complexity
- The cross-validation sample may share characteristics with the development sample and further cross-validation is intended prior to application

The Development of a Dynamic Progress Feedback System to Guide Psychological Treatment in Primary Care

In England, psychological therapy is often accessed via Improving Access to Psychological Therapies (IAPT) services, which deliver evidence-based psychological interventions organised in a stepped care model (Clark, 2011). Between 2016-2017, 50.8% of people attending IAPT were classified as recovered according to standardised depression and anxiety measures (NHS Digital, 2018). This figure is lower when considering the criteria of 'reliable and clinically significant improvement' (RCSI), which defines recovery as both change that is not down to measurement error and scores that move to below clinical cut-offs (Jacobson & Truax, 1991). Lutz (2002) highlighted the issue that although we have evidence of the effectiveness of psychological interventions, we also need patient-focused research to understand why not everyone responds.

A key goal of patient-focused research is to enable clinicians to identify clients that may be at risk of poor progress so that they can adjust therapy in a timely way to improve outcomes. Clinical judgment is considered to be limited as a means of deciding whether patients are making progress in treatment, being biased towards over-optimism (Hannan et al., 2005). It is therefore necessary for clinicians to back judgments up with rational and empirical feedback methods (Lutz, Stultz, Martinovich, Leon, & Saunders, 2009). Research indicates that clients who are not responding to therapy can be identified using progress feedback methods, and that having this feedback can improve outcomes (Delgadillo et al., 2018; Lambert, Whipple, Kleinstäuber, Hilsenroth, & Norcross, 2018). There are currently two standard approaches to identify cases at risk of poor outcomes: patient profiling and outcome feedback models.

Profiling and Feedback Methods

Patient profiling models make a prediction at the outset of therapy based on baseline symptom scores or other key variables. Several profiling methods have been proposed, including the nearest neighbours model (Lutz et al., 2005), latent class analysis (Saunders, Cape, Fearon, & Pilling, 2016), or risk stratification models (Delgadillo, Huey, Bennett, & McMillan, 2017). For example, the Leeds Risk Index ([LRI] Delgadillo, Moreea, & Lutz, 2016) is a patient profiling tool that expanded on previous models through combining key demographic variables into a weighted risk score (range 0-21), classifying people as being at high, moderate or low risk of poor treatment outcomes at the beginning of therapy.

Outcome feedback systems use session-by-session outcomes data to statistically compare an individual's scores against group-based normative values, thus enabling the identification of clients whose symptoms are significantly more severe than expected (Lutz et al., 2009). Some feedback models are based on monitoring whether clients' symptoms conform to "expected treatment response" (ETR) curves. The ETR concept is based on the dose-response model, where the assumption of a negatively accelerating relationship between improvement in outcomes and number of sessions received is made (Howard, Kopta, Krause, & Orlinksy, 1986; Lutz, 2002). When an individual's symptom scores fall outside of the ETR curve boundaries, a risk signal is created to alert the clinician that someone is "not on track". Following this feedback, clinicians are prompted to identify and address potential problems that may be interfering with effective treatment.

Different Patterns of Recovery

The ability of these models to accurately identify clients who are "not on track" is essential to prevent treatment failure. However, whilst the above approaches have

furthered our abilities to identify patients at risk, they rely on static snapshots of a person's scores and may not fully account for different patterns of treatment response. For example, there have been challenges to the idea that everyone follows an average dose-response pattern of recovery. As discussed in section one, the "Good-enough level" (GEL) perspective argues that dose-response curves may be an artefact of the aggregation of different samples at different timepoints, with harder to treat problems requiring therapy for longer durations, thus creating the illusion of a diminishing effect over time.

Recent studies have examined such differential patterns of change, finding for example that early sudden gains or losses are important indicators of progress (Delgadillo et al., 2014; Lutz et al., 2012; Rubel, Lutz, & Schulte, 2015). Other studies have noted that different problems may respond at different rates (Barkham et al., 1996; Owen, Adelson, Budge, Kopta, & Reese, 2016), and that some people may leave treatment early due to making progress on 'easier' symptoms such as wellbeing, without making progress on slower to respond difficulties such as life functioning or relationships (Kivlighan, Lin, Egan, Pickett, & Goldberg, 2018).

Research Rationale and Aims

Not all recovery trajectories therefore fit a negatively accelerating or ETR curve, meaning that there are subsets of people who are not accurately monitored using standard, static, techniques. It is of interest, therefore, to examine whether different techniques could improve our ability to identify and support clients who are not on track. The field of predictive modelling is relatively young in mental health settings. However, it is an area of importance, where clinicians who are better supported to accurately and rapidly identify cases at risk can respond accordingly to improve outcomes. For example, through having better understanding of who is "not on track",

the clinician is able to make better use of supervision to reformulate and to assess interventions, manage ruptures, or to consider other necessary sources of support.

For these reasons, the current study aimed to develop dynamic progress feedback models, combining the initial profiling capability of the LRI (Delgadillo et al., 2016) with sessional progress information to create personalised prognoses of recovery. The models would be dynamic in that they would 'learn' from past information as well as incoming progress scores, recalculating recovery prognoses to provide updated percentage probabilities of seeing RCSI. In the interests of parsimony, or the preference for simplicity over unnecessary complexity, it was also considered important to compare complex dynamic models with simpler dynamic and static models. If such models were considered to be valid, there would be the potential for them to be developed into a software application and used in psychotherapy settings in the UK as a clinician support tool. This would not only have the potential to improve outcomes for individuals but could save mental health services time and money, where re-referrals are otherwise likely.

Research Questions

The research questions (RQ), aims (A), and objectives (OB) are numbered and correspond with each other for cross-referencing.

RQ1. Can we integrate patient profile information and routine progress monitoring into dynamic prediction systems? (A1, OB1, OB2).

RQ2. Do these integrated prediction systems generalise to another sample? (A2, OB4, OB5).

RQ3. Do complex models outperform parsimonious ones? (A3, OB3, OB6).

Aims

A1. The first aim was to use a pre-existing database of patient outcomes from IAPT Leeds (dataset one) to develop a dynamic progress feedback system. A sub-aim was to build models of increasing complexity to allow for comparisons between basic and more complex feedback systems (RQ1, OB1, OB2).

A2. The second aim was to cross-validate these models using a new dataset from IAPT Cumbria (dataset two) as a measure of generalisability (RQ2, OB4, OB5).

A3. The third aim was to examine whether more complex models outperform basic models in both datasets (RQ3, OB3, OB6).

Method

Design

This was a retrospective database analysis of clinical case records involving two phases: model development in the first phase, and cross-validation in the second phase (figure 2).

Setting and Sample

Setting. Data were provided by Leeds and Cumbria NHS IAPT services between 2016-2018. The National Institute for Health and Care Excellence recommends a stepped care model for mild-moderate cases of anxiety or depression, where patients are matched to the least restrictive step first (step two low intensity), whilst also ensuring that they receive the optimum care (e.g. going straight to step three high intensity if this is clinically indicated) (NICE, 2011). Low intensity interventions (LIT) are usually based on Cognitive Behaviour Therapy (CBT) in a guided self-help or group format. High intensity interventions (HIT) are often CBT-based and delivered by high intensity therapists for people with moderate-severe symptoms (NHS England, 2015).

Sample. Research suggests that at least three timepoints are necessary to capture non-linear trends, with an additional timepoint necessary to ensure outcomes are not

confounded with predictor variables (Rubel et al., 2015). At least four therapy sessions have been shown as necessary to achieve more than 50% RCSI rates (Delgadillo et al., 2014). Samples having at least four sessions were therefore included in the analyses. Overall, *n*=2494 cases from Leeds and *n*=2084 cases from Cumbria were included, however sample sizes per session varied (see results/appendix A). Data from IAPT Leeds were pre-existing, however NHS ethical approval was granted for IAPT Cumbria to provide the new cross-validation dataset (appendix B).

Samples across the services differed in several ways particularly at step two: Leeds used more group interventions than Cumbria, had a higher percentage of high LRI groups, and saw lower RCSI rates. Differences were also observed between included and excluded samples. Mean baseline scores and RCSI rates were significantly higher for PHQ-9 and GAD-7 in the included samples, which is to be expected given the inclusion criteria. Mean age was significantly lower in the included sample at Step3 for Leeds, and significantly higher in the included sample at Step 2 in Cumbria. Significantly more females than males were also in the included versus excluded sample at Step 3 in Leeds (see appendix A for all figures). The following inclusion criteria applied:

Table 1

Participant Inclusion Criteria

Clients aged 16 and above (data from Cumbria included 18 and above).
Clients identified as having clinically significant anxiety or depression scores as evidenced by their PHQ-9 and GAD-7 scores.
Clients with a minimum of four therapy sessions.
Clients received low and/or high intensity interventions within IAPT.

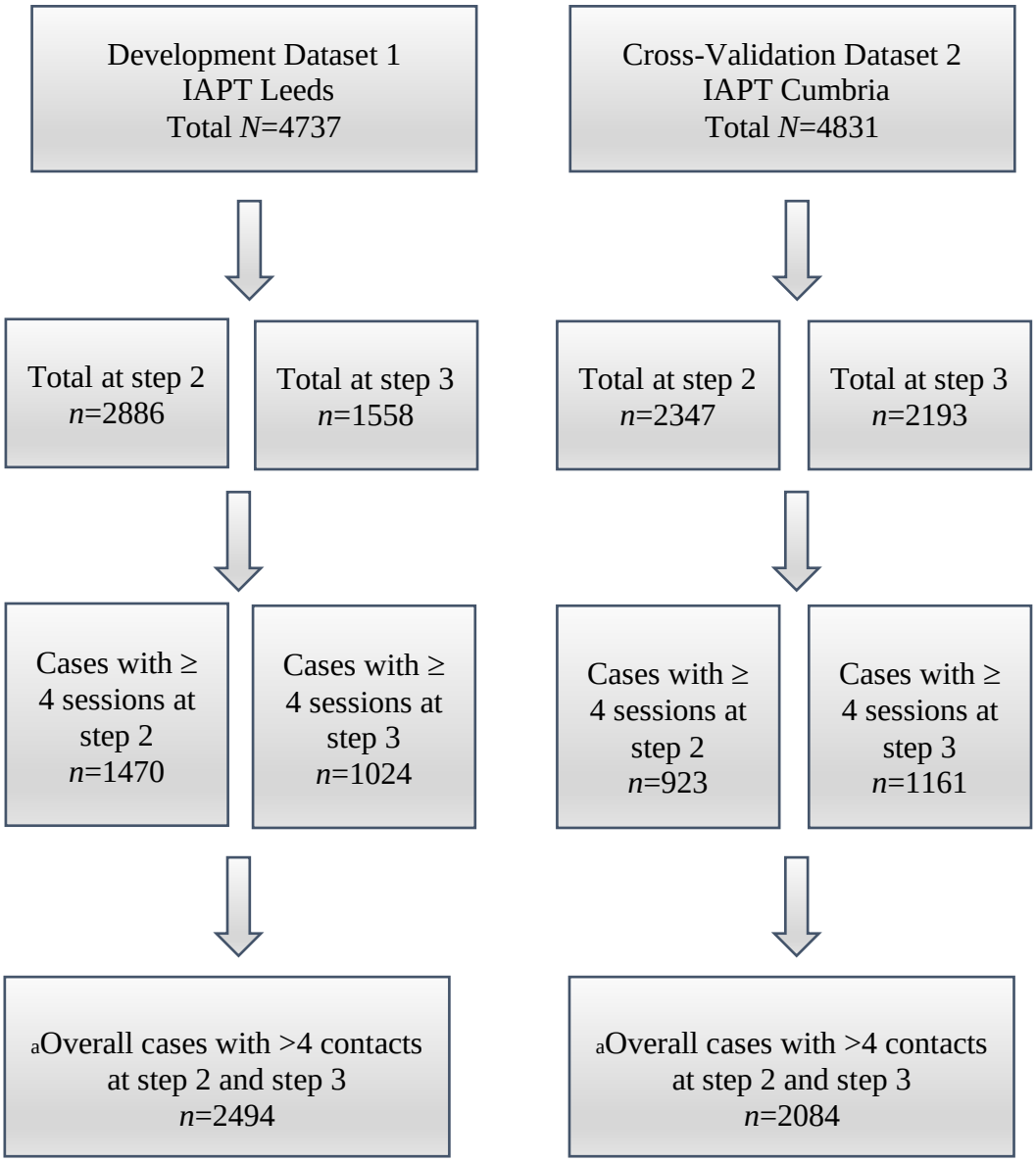


Figure 1. Overall sample numbers in datasets one and two with valid PHQ-9 or GAD-7 RCSI figures (who began treatment above cut-off). Precise sample figures per model are provided in appendix A.

aNote that some cases received therapy at step 2 and step 3.

Table 2

Sample Demographics for Included Cases (Minimum Four Sessions and Valid PHQ-9 or GAD-7 RCSI Rating) and Baseline Score Comparisons Between Included and Excluded Samples

	Dataset 1: IAPT Leeds		Dataset 2: IAPT Cumbria	
	Step 2	Step 3	Step 2	Step 3
Mean age (range)	37 (16-88)	36 (16-81)	42 (18-87)	41 (18-86)
Gender	64.8% Female 35.2% Male	69.3% Female 30.7% Male	64.2% Female 35.8% Male	61% Female 39% Male
[a] Ethnicity	85.6% White, 10.6% Other, 3.8% Other White	88.1% White, 8.7% Other, 3.2% Other White	95.7% White, 2.1% Other, 2.3% Other White	96.7% White, 2% Other White, 1.3% Other
Employment	67.3% Employed	62% Employed	77.6% Employed	68% Employed
Initial step	68.1%	31.8%	53.6%	46.4%
Main interventions	97.9% Groups, 2% CBT, .1% EMDR	77.4% CBT, 14.4% counselling, 5.9% Groups, 2.2% EMDR	85.9% Self-help, 11.4% Counselling, 1.9% Groups, 0.9% CBT	86.9% CBT, 7.6% Therapy other, 2.9% Groups, 1.9% Counselling, .7% Self-help
[b] Main primary presenting problems	27.3% Depression, 46% mixed anxiety & depression, 21.8% anxiety, 2.3% other	29% Depression, 34.3% mixed anxiety and depression, 21.5% anxiety, 6.7% OCD, 4.2% trauma	27.5% Depression, 58.8% anxiety, 11% other	45.6% Depression, 35.7% anxiety, 11.4% trauma

[c]Mean baselines (SD and range) and rate of RCSI in included sample per measure	PHQ-9: 16.22 (4.4, 10-27). RCSI: 40.8% GAD-7: 14.29 (3.801, 8-21) RCSI: 42.1% WSAS: 17.61 (8.771, 0-40)	PHQ-9: 16.78 (4.432, 10-27). RCSI: 50.9% GAD-7: 14.82 (3.752, 8-21) RCSI: 50.9% WSAS: 19.60 (8.885, 0-40)	PHQ-9: 15.44 (3.941, 10-27). RCSI: 65.6% GAD-7: 13.97 (3.731, 8-21) RCSI: 65.8% WSAS: 19.22 (9.45, 0-40)	PHQ-9: 17.22 (4.496, 10-27) RCSI: 60.5% GAD-7: 15.24 (3.924, 1-21) RCSI: 56.8% WSAS: 23.46 (8.697, 0-40)
[d]Mean baselines (SD and range) and RCSI rates in excluded sample per measure	PHQ-9: 14.48 (6.244, 0-27). RCSI: 8.7% GAD-7: 13.39 (5.181, 0-21) RCSI: 12.2% WSAS: 19.70 (9.066, 0-40)	PHQ-9: 14.50 (6.580, 0-27). RCSI: 10.9% GAD-7: 13.38 (5.588, 1-21) RCSI: 12.2% WSAS: 20.18 (9.540, 0-40)	PHQ-9: 11.81 (6.290, 0-27) RCSI 23.3% GAD-7: 11.31 (5.564, 0-21) RCSI 21.4% WSAS: 18.87 (9.434, 0-40).	PHQ-9: 14.81 (6.479, 0-27) RCSI 15% GAD-7: 13.30 (5.670, 0-21) RCSI 14.2% WSAS: 23.64 (8.989, 0-40).
LRI groups included sample	Low 41.1%, Mod 44.2%, High 14.7%	Low 29.9%, Mod 49.2%, High 20.9%	Low 39%, Mod 51%, High 10.1%	Low 18.2%, Mod 55.6% High 26.2%
LRI groups excluded sample	Low 31.6%, Mod 47.7%, High 20.7%	Low 21.8%, Mod 47.6%, High 30.5%	Low 36.2%, Mod 51.6%, High 12.2%	Low 15.3%, Mod 54.2%, High 30.5%

[a]Ethnicity categories reduced to prevent identification of participants where numbers were low.

[b]As recorded in clinical notes.

[c]RCSI rates based on included sample with minimum four sessions and valid RCSI figures (e.g. began treatment above cut-off).

[d]RCSI rates based on excluded sample (e.g. including those who had <4 sessions and began treatment before cut-off).

Sample size calculation. Sample size calculations for logistic regression were guided by Delgadillo et al.'s (2014) research on outcomes prediction, Delgadillo et al.'s (2016) LRI study, and Hsieh's (1989) sample size tables for logistic regression (appendix C). The supplementation of initial profile scores with session scores was expected to yield more robust predictive models than using initial profile scores alone, hence the diagnostic odds ratios from Delgadillo et al.'s (2014) research were considered appropriate for the calculation. Hsieh recommends Whittemore's formula for cases where the dependent variable is dichotomous (such as RCSI) and where risk factors are continuous and have a joint multivariate normal distribution. The calculation is based on event rate, an odds ratio, alpha (0.5 as the acceptable level of significance), and power (set at 90%). The event rate is taken from Delgadillo et al.'s (2016) LRI study (46%), and the odds ratio is taken from Delgadillo et al.'s (2014) outcomes prediction study (2.10):

Table 3

Sample Size Calculations

Event rate: RCSI = 46%
Odds ratio: β = 2.10
Alpha: = 5%
1 – β: 90% (power)
Using Hsieh's (1989) Table III (for one covariate): P(0.45) and r(2.10): N = 99
Formula adjusted for multiple logistic regression (the conservative estimate of correlation coefficient is 0.5, guided by Delgadillo et al., 2014): 99 / (1 – (0.5*0.5)) = 99 / (1 – 0.25) = 99 / 0.75 = 132.

Data from *N*=132 participants were therefore needed to construct the model and the same to cross-validate it.

Data Security and Anonymity

Datasets were stored in a secure University drive which was only accessible to the researchers. The Leeds dataset was pre-existing, however the Cumbria data files were extracted by IAPT Cumbria and transferred using a secure Dropbox facility with a temporary password (using NHS secure file transfer). The data remains the property of Leeds and Cumbria NHS Foundation Trusts and any mishandling or loss of data would be subject to professional liability procedures as dictated by the trusts. Anonymity was maintained throughout the study, and individual cases were only identifiable through a number assigned by the electronic database. Therefore, it was not possible for any researchers to personally identify any patient.

Outcome Measures

The Patient Health Questionnare-9 ([PHQ-9), Kroenke, Spitzer, & Williams, 2001) and the Generalised Anxiety Disorder-7 ([GAD-7], Spitzer, Kroenke, Williams, & Lowe, 2006) are both routinely used in IAPT services. The LRI (Delgadillo et al., 2016) was introduced to both participating IAPT services in 2016 (appendix D).

i) The PHQ-9 is a nine-item self-report Likert questionnaire assessing depression symptom severity. Scores range from 0-27, with clinical caseness being scores of 10 or over. The PHQ-9 was found to have good internal reliability (Cronbach's alpha .89 and .86), excellent test re-test reliability, and sensitivity and specificity of 88% for scores of 10 or more.

ii) The GAD-7 is a seven-item self-report Likert questionnaire assessing anxiety symptom severity. Scores range from 0-21, with clinical

caseness being scores of 8 or over. Good internal consistency was found (Cronbach's alpha .92), and test-retest reliability (.83), with sensitivity of 92% and specificity of 76% for scores of 8 or more

iii) The LRI is a risk profiling tool which uses the factors younger age, unemployment, disability, impaired functioning, low expectancy of therapy and depression symptom severity to provide a weighted risk score, classifying people into low (<=4), moderate (5-9) or high (>10) risk of not achieving RCSI. High LRI scores were shown to be significantly correlated with cases who were "not on track" (using ETR boundaries throughout therapy), and rates of RCSI were found to be significantly lower in those who had moderate and high LRI scores.

Operational Definition of Primary Outcome

Adequate outcomes are defined here as those who attain RCSI by the end of treatment, referring to both reliable *and* clinically significant change. Reliable change is calculated to ascertain whether change is down to measurement error, which is equal to a patient's pre-post treatment change score divided by the standard error of difference (Jacobson & Truax, 1991). Richards and Borgin (2011) used this method to identify reliable change scores in a sample from IAPT similar to the current study, calculating this to be 5 for GAD-7 and 6 for PHQ-9. Clinically significant change refers to a person's scores moving from the clinical range into the non-clinical range by the end of treatment (Evans, Margison, & Barkham, 1998). The clinical cut-off scores for the PHQ-9 and GAD-7 are suggested as 10 and 8 respectively (Kroenke et al., 2001; Spitzer et al., 2006). There were therefore four criteria necessary to establish RCSI in this study:

Table 4

Criteria to Establish RCSI

i	Cases must have at least four measures (in order to reasonably calculate RCSI during the first month of treatment).
ii	Pre-treatment scores must be in the clinical range (10 or over in PHQ-9 and 8 or over in GAD-7).
iii	Post-treatment scores must be in the non-clinical range.
iv	Scores must have improved by at least 6 points for the PHQ-9 or 5 points for the GAD-7.

Analysis

Data cleaning and preparation. The Statistical Package for the Social Sciences ([SPSS] version 25) software was used for all analyses. Initial data cleaning and preparation took place both datasets. Databases were inspected for duplicate entries, missing data values or inputting errors, and data ranges checked. Key variables were coded identically. Missing data were treated as 'missing at random' and were not imputed as this was not considered problematic in a large dataset. Only contacts attended were included in the analyses.

Model phases. The analysis consisted of two phases: model development in dataset one and cross-validation in dataset two:

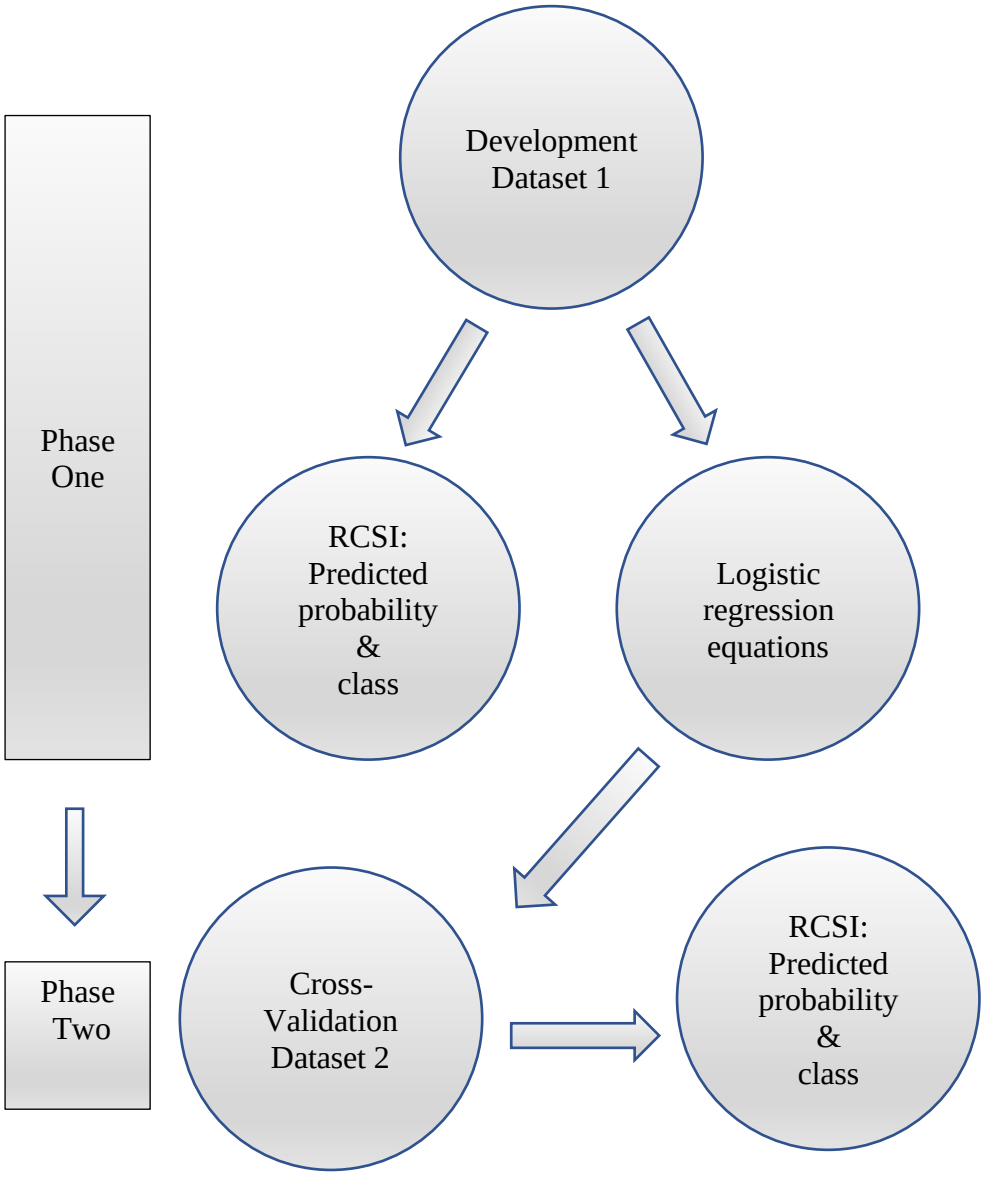


Figure 2. Two phases and outputs: model development in dataset 1 and cross-validation in dataset 2.

Phase One: Model Development in Dataset One

Objective 1 (A1, RQ1). **Identifying and creating key variables**. LRI scores at assessment were used to group cases into low, moderate or high LRI classes, which were used as categorical predictors in the models (Delgadillo et al., 2016). The LRI includes a PHQ-9 element therefore it also provided a measure of symptom severity

from assessment to session one. Baseline scores (at session one of therapy) and subsequent session scores on the PHQ-9 and GAD-7 were also used as predictors in logistic regression models. Two further variables were created from the existing session data: Risk Sum (RS) and Standard Deviation (SD). These variables were designed to capture cumulative information from previous sessions:

Table 5

Variable Descriptions

Variable	Description
1. LRI Profile (LRI)	LRI profile group taken at assessment.
2. Baseline (BL)	Scores on PHQ-9 and GAD-7 taken at session one of therapy.
3. Session score (SS)	Subsequent session scores on PHQ-9 and GAD-7.
4. Risk sum (RS)	A sample level predictor, where a cumulative risk score was calculated each time an individual's score exceeded the sample mean plus one standard deviation.
5. Standard deviation (SD)	An individual level predictor, where the individual's own standard deviation is summed cumulatively from session to session, where greater symptom variability would be expected with greater change (Shalom et al., 2018).

Objective 2 (A1, RQ1). **Model development in dataset one.** Logistic regression was used to create models of increasing complexity, based on predicting the binary outcome of seeing RCSI or not. In the dynamic models (involving more than one predictor), algorithms were created that combined prior probabilities from the intercepts (baseline scores or LRI groups) with new information from the slopes (weekly progress scores and computations of these) to calculate a posterior probability distribution for RCSI. Each new weekly progress score would result in a re-estimation of the probability distribution, translating to a percentage probability of attaining RCSI.

Predicted class outputs were also calculated and coded as 1=RCSI and 0=No RCSI, where RCSI was classed as positive if the percentage probability value was >=0.5 or 50%.

Box 1. Case example

Take a hypothetical case example "Tom" at session three HIT.

The question the models seek to address is this:

"Compared with the rest of the sample at session three HIT, what is the probability of someone with Tom's LRI group, baseline score, session three score, and cumulative risk sum and standard deviation, seeing RCSI at his final session of treatment?"

If Tom's probability output was 0.637 at session three, this would mean that his predicted classification for the end of treatment would be 1=RCSI, and he would have about a 64% chance of achieving this.

Modelling Strategy. Four modelling streams were identified (table 6). Participant data was separated into LIT and HIT treatment (step two and step three). It was expected that there may be differences between those accessing low intensity versus high intensity treatment, which would otherwise be masked. Some cases would have stepped up from low to high intensity treatment, however it was considered pragmatic to analyse these as separate incidents of therapy rather than analysing these cases as a sub-group. Where participants did step up to HIT, the first session at HIT would therefore be classed as session one. PHQ-9 and GAD-7 outcomes were also

examined separately in order to accommodate any differences in outcomes between anxiety and depression symptoms.

Table 6

Four modelling streams

Stream A	Low intensity (LIT) PHQ-9 (up to 8 sessions)
Stream B	Low intensity (LIT) GAD-7 (up to 8 sessions)
Stream C	High intensity (HIT) PHQ-9 (up to 12 sessions)
Stream D	High intensity (HIT) GAD-7 (up to 12 sessions)

Models were built for eight therapy sessions at LIT (median N=5, range 1-24), and for 12 at HIT (median N=12, range 1-30). Some cases had more than 12 sessions at HIT, however the majority of therapy reviews would have taken place by session 12, and the value of these predictive models would be in alerting clinicians at earlier stages. Sample numbers were also too low at later stages to have adequate power and reliability. Six explanatory models were constructed per session within the four streams (table 7, figure 3). RCSI was calculated based on the final session of therapy at low or high intensity, and selection rules applied to each model per session in line with this.

As discussed, a minimum of four sessions overall was considered necessary to achieve at least 50% RCSI rates (Delgadillo et al., 2014). When selecting cases for models at sessions one to three, therefore, a selection rule applied that they must have >=four sessions of therapy in total. For subsequent models, the minimum total number of sessions required would always be one more than the current session number so that a prediction of RCSI was possible. For example, modelling outcomes based on scores at session six would require that cases had at least seven sessions in total, and RCSI would be based on that individual's final therapy session. The resultant models

ran from static single predictors to combinations of increasing complexity, in order to assess whether more complex models perform better. The LRI was always added last to see if it significantly added value beyond the other variables, given that this was the more resource heavy measure for services to use.

Chi-square analyses indicated whether individual variables significantly added value to the models, and backward elimination was used until only significant variables remained in the final predictive models. There were two instances where non-significant variables may have been retained. Baseline was always retained as a measure of change, and overall patterns of significance were identified in the data and adhered to. For example, if a variable was significant at session three, but became borderline at session four before becoming significant again at session five onwards, it would be retained because there may be session sample anomalies.

Table 7

Model Descriptions

Model	Variables examined
1. LRI Profile	LRI group as defined at assessment used as a single static predictor.
2. Session score	PHQ-9 and GAD-7 session scores as single predictors.
3. Basic dynamic	Baseline PHQ-9 or GAD-7 scores plus current session score.
4. LRI dynamic	Baseline PHQ-9 or GAD-7, current session score, plus LRI.
5. Progressive dynamic	Uses all variables, apart from the LRI (BL, SS, RS, SD).
6. Complex dynamic	Uses all five variables (BL, SS, RS, SD, LRI).

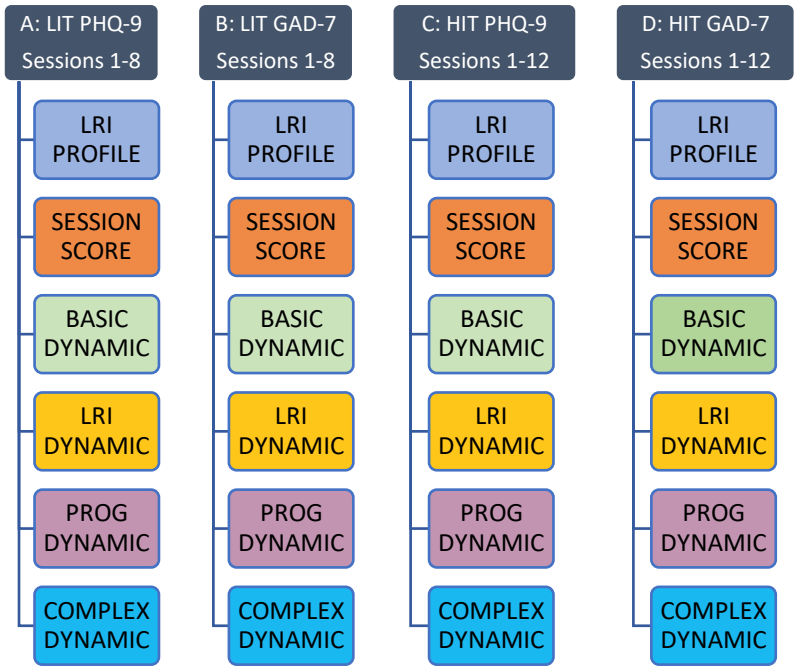


Figure 3. Modelling streams and models

Note: Selection rules applied per session as described above. RCSI was based on final outcome measure per step.

Objective 3: Comparing complex versus basic models (RQ3, A3). Models were compared on how much of the variation (Nagelkerke R_2) in outcomes (RCSI) they were capable of explaining. Classification tables also indicated the percentage accuracy of each model in making RCSI classifications, and are included in appendix A.

Objective 4: Preparing data for cross-validation (RQ2, A2). Logistic regression equations were written using the ß-coefficients from the outputs, which represent 'coordinates' for the slope and intercept for each predictor. For example:

$$p = \exp(ß_0 + ß_1x_1 + ß_2x_2 + \ldots + ß_qx_q) / 1+\exp(ß_0 + ß_1x_1 + ß_2x_2 + \ldots + ß_qx_q)$$

…where p is the predicted probability of seeing RCSI, $ß_0$ is the constant. and $ß_1$ etc. are the ß-coefficients of predictor variables, multiplied by the corresponding 'incoming' variable values (x_1). For example, the ß-coefficient of low intensity baseline PHQ-9, multiplied by the actual baseline PHQ-9 score, and so on.

New probability variables were then created in cross-validation dataset two using these equations. Sensitivity and specificity analyses were performed using Receiver Operating Characteristic (ROC) curve coordinates, to assess whether using a balance between sensitivity and specificity would yield better classification results than the default cut off of 0.5 (50%) for RCSI. Kappa analyses indicated that 0.5 was generally a superior cut-off, however, so this value was used to create new predicted classification variables in dataset two (>=0.5=RCSI). These new predicted probability and class variables in dataset two were then ready for cross-validation analyses.

Phase two: Cross-Validation in Dataset Two (Cumbria)

Objective 5: Kappa, Area under the Receiver Operating Characteristics Curve (AUC) and Brier scores (A2, RQ2). Cross-validation took place on *n*=2084 available cases from IAPT Cumbria, although sample sizes varied per session (see results figures). Three methods were used to assess how well the predictions performed in the new sample:

i) AUC analyses were used to evaluate classification performance, assessing the rate of true and false positives and negatives based on the models' predicted probabilities and observed RCSIs. AUC provides a figure between 0-1, where 0.5 is no better than chance and 1 is perfect classification (Hosmer & Lemeshow, 2013).

ii) Kappa analyses were also used to assess absolute agreement between the predicted and observed RCSI classifications, where 1=perfect agreement, 0=agreement by chance, and -1=perfect disagreement (Watson & Petrie, 2010).

iii) Brier scores assess the error of a probability forecast where 0=complete accuracy and 1=complete inaccuracy, incorporating domains of reliability, resolution and uncertainty (Brier, 1950; Rufibach, 2010; Tetlock & Gardner, 2016). Brier scores were calculated on an individual basis and averaged to provide an overall session by session Brier score for each model, using the mean squared error formula (Redelmeier, Block & Hickam, 1991):

$$\text{Brier} = \frac{1}{N} \Sigma \, (p - o)^2$$

…where the difference between predicted probability (p) and observed RCSI (o) was calculated for each person and squared, these squared values were summed (Σ), and divided by the sample number (N) to find the mean score.

Objective 6: Dataset two model comparisons (RQ3, A3). Comparisons were made between the models via graphing their AUC, Kappa and Brier figures and using confidence intervals, where overlap would indicate differences were not significant. Kappa does not provide confidence intervals so these were calculated based on standard error figures. It is not possible to calculate confidence intervals for Brier scores, however the three methods were used to supplement each other to inform decisions on model superiority.

Ethical Implications

Data were provided by people who consented for it to be used anonymously for the purposes of research when they began treatment in IAPT. These participants were provided with patient information leaflets containing contact details for further information, queries, or to withdraw their consent (appendix E).

As the study used pre-existing anonymous data and did not constitute any form of interference with patients an NHS proportionate review of ethics was granted. The

aim of the study was to improve clinicians' ability to detect patients who are not on track, hence the possibility of harm being caused due to the study was considered minimal. Possible concerns might be that knowing that someone is not on track could engender hopelessness. However, research suggests that providing feedback improves patient outcomes (Delgadillo et al., 2018). Furthermore, feedback measures are already routinely used in practice, so this study only seeks to improve the accuracy of a system that is already in place.

Patient and Public Involvement

A patient representative was involved in discussions at IAPT Cumbria about providing data for this study, and was in support. Subsequent research would look at developing a user-friendly software interface for clinicians in IAPT, where consultation would be key.

Dissemination

This study will be prepared for publication in a scientific journal and is intended for future development into a clinical tool.

Results

Phase One: Model Development in Dataset One (A1, RQ1, OB1-2).

The following tables and graphs present variance explained (Nagelkerke R_2) for each model per session in the different streams (see appendix A for full results tables). There are no firm guidelines for acceptable interpretations of Nagelkerke R_2, however therapist effects have been shown to explain between 1-10% of the variance (Saxon & Barkham, 2012). Lower figures would generally be expected when making more distal predictions in psychotherapy than proximal and tightly controlled settings. As discussed, predictors were only retained if they were significant (apart from the circumstances described), and are detailed below the figures. Note that complex models

revert to being equivalent to more basic ones at some sessions, due to the backward elimination of non-significant predictors. The sample number required to build and test the models was *N*=132. Sample numbers were the same for all models apart from in the LRI profile only models, which were marginally higher due to fewer selection criteria; the lower figure is therefore reported in tables below. Sample numbers were too low (*n*=97) at session eight (e.g. S8) at LIT for reliable interpretation, but figures are presented for information.

Model Stream A: LIT PHQ-9

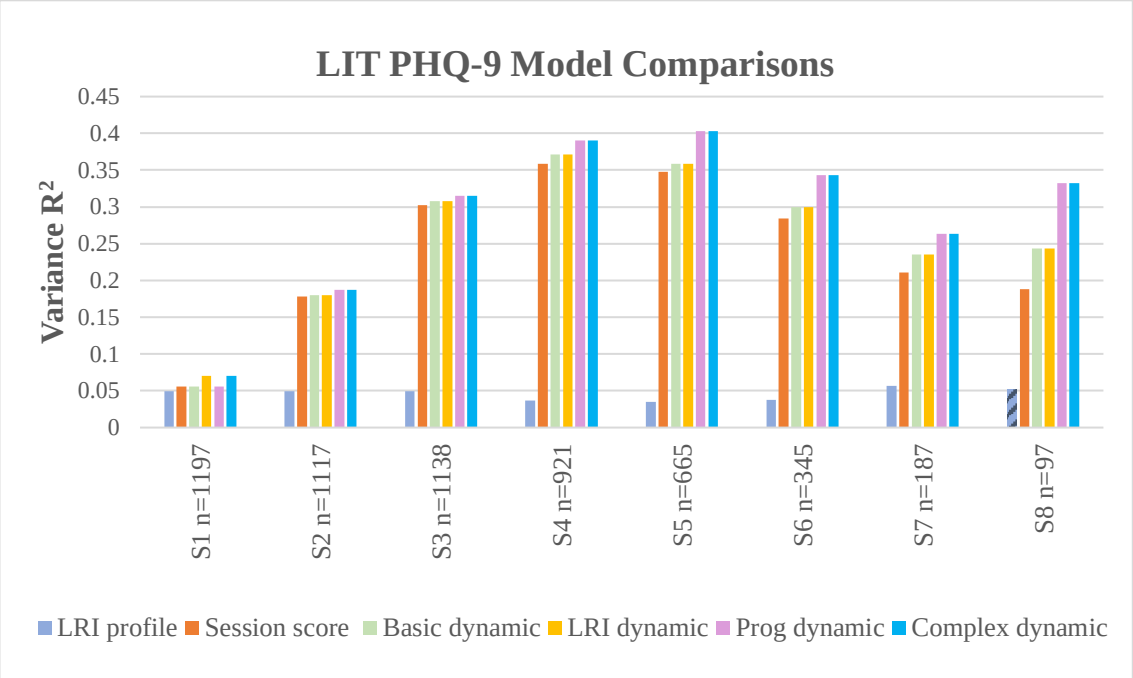


Figure 4. LIT PHQ-9 Comparisons of Variance Explained (Nagelkerke $R2$).

Notes
All models are significant at the p<0.001 level apart from LRI profile only p=<.01 at S6, p=<.05 at S7, and non-significant at S8.
LRI dynamic includes BL and SS at every session, however the LRI was only included at S1, making it equivalent to the basic dynamic model thereafter.
Progressive dynamic includes all variables apart from the LRI, with RS significant from S2-S6 and SD from S4-S8.
Complex dynamic includes all variables, with RS from S2-S6 and SD from S4-S8. The LRI was only significant at session 1, making it equivalent to the progressive dynamic model thereafter.

Model Stream B: LIT GAD-7

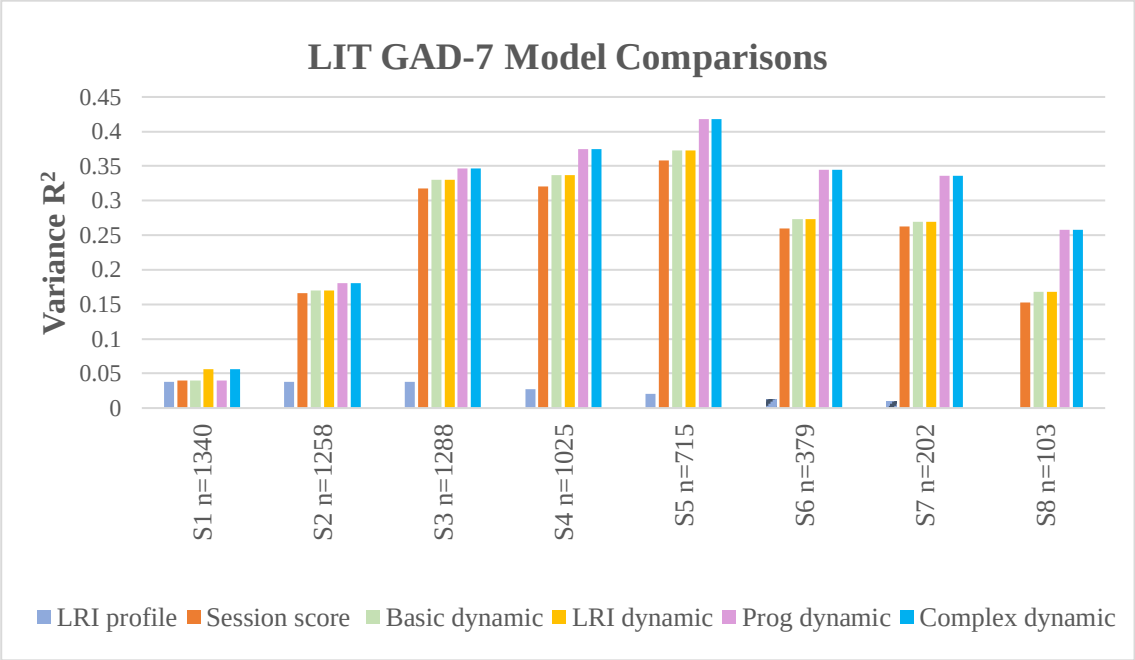


Figure 5. LIT GAD-7 Comparisons of Variance Explained (Nagelkerke $R2$).

Notes
All models are significant at the $p<0.001$ level apart from LRI profile only $p<.01$ at S5, and non-significant from S6-S8, and Session score only model $p<.01$ at S8. LRI dynamic includes BL and SS at every session, however the LRI was only included at S1, making it equivalent to the basic dynamic model. Progressive dynamic includes all variables apart from the LRI, with RS significant from S2-S7 and SD from S3-S8. Complex dynamic includes all variables, with RS from S2-S6 and SD from S3-S8. The LRI was only significant at session 1, making the model equivalent to the progressive dynamic model thereafter.

Model Stream C: HIT PHQ-9

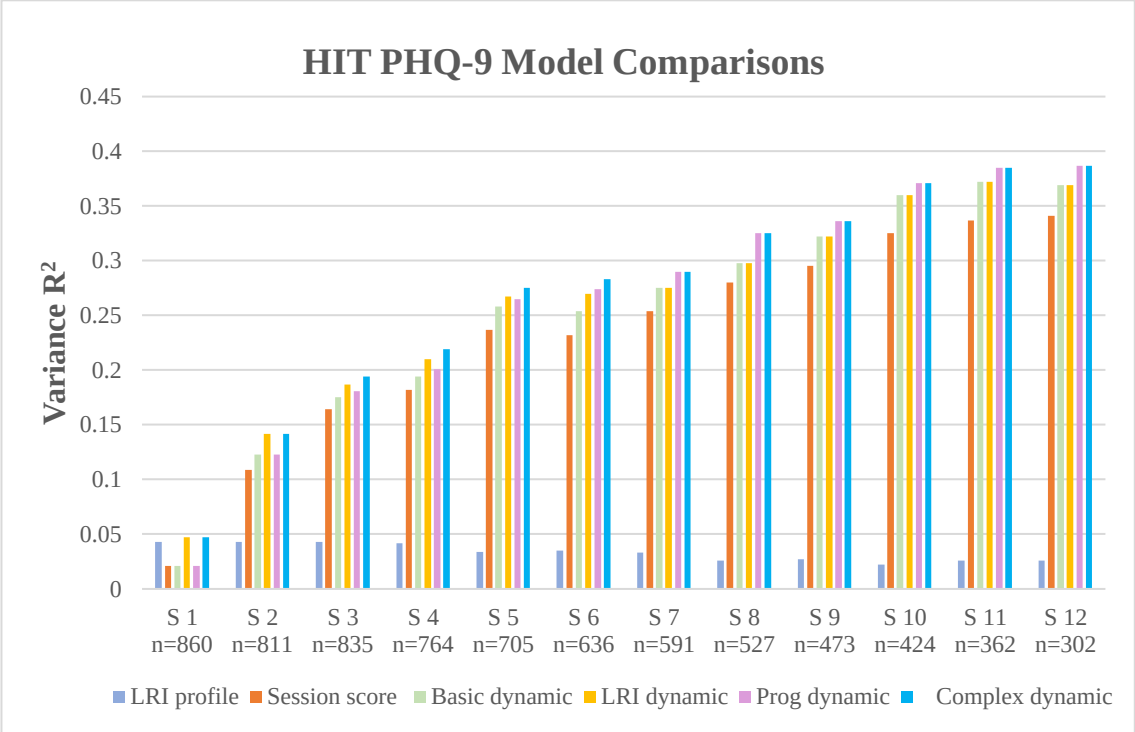


Figure 6. HIT PHQ-9 Comparisons of Variance Explained (Nagelkerke R^2).

Notes
All models are significant at the p<0.001 level, apart from LRI profile only $p<.01$ at S8-S9, and $p<.05$ at S10-S12. LRI dynamic includes BL and SS at every session, and the LRI from S1-S6. Progressive dynamic includes all variables apart from the LRI, with RS significant from S6-S12 and SD from S3-S8. Complex dynamic includes all variables, with RS from S7-S12, SD from S3-S8, and the LRI from S1-S6.

Model Stream D: HIT GAD-7

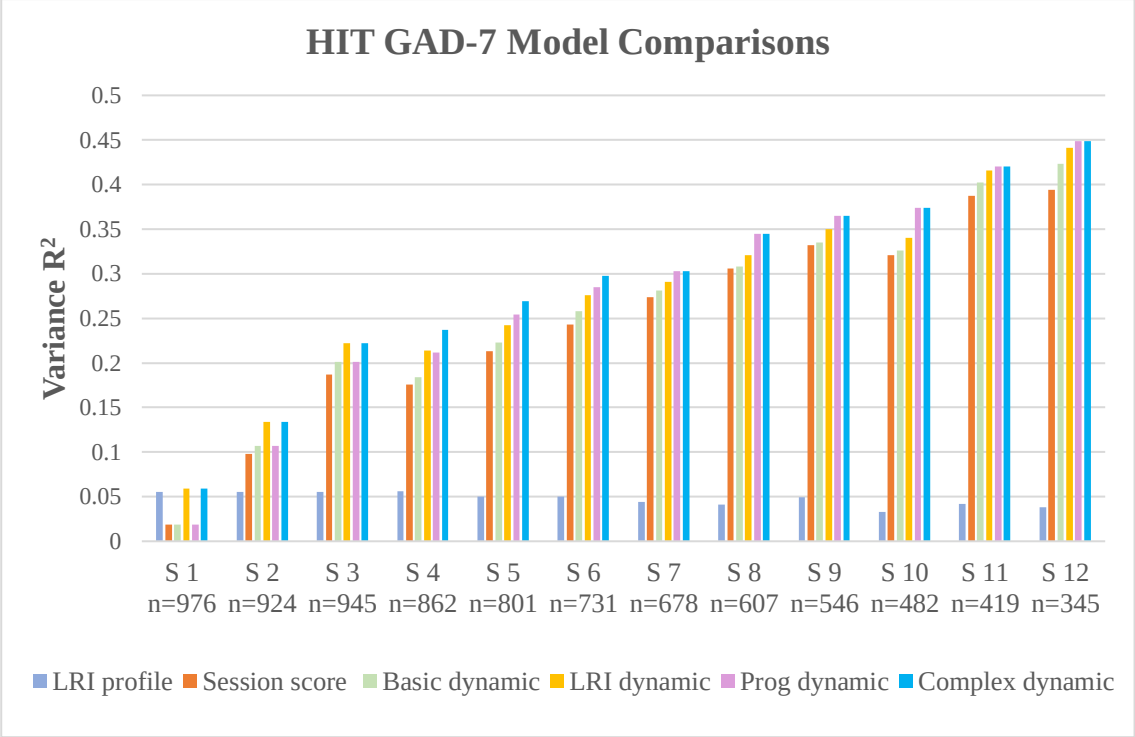


Figure 7. HIT GAD-7 Comparisons of Variance Explained (Nagelkerke R_2).

Notes
All models are significant at the p<0.001 level, apart from LRI profile only, p<.01 at S10 and S12, and LRI dynamic p<.01 at S6 and S8 and p<.05 at S7, and S9-12. LRI dynamic includes BL, SS and LRI at all sessions. Progressive dynamic includes all variables apart from the LRI, with RS significant from S4-S12 and SD from S4-S10. Complex dynamic includes all variables, with RS from S4-S12 and SD from S4-S10, and the LRI from S1-S6.

Summary of Model Development

(A1, RQ1, OB1-2)

Low intensity. All models showed significant ability to predict outcomes apart from the LRI profile model at latter sessions. In the dynamic models, the LRI did not add value beyond session one, meaning that the complex model was equivalent to the progressive beyond session one. It is desirable for clinicians to know whether someone is on track as early as possible, particularly where research suggests that 31% may drop out of treatment by session four (Delgadillo et al., 2014). At sessions three and four on the PHQ-9, the Progressive model explained the most variance: 32% ($\chi_2(3) = 303.263$, $p < .001$) and 39% ($\chi_2(4) = 316.626$, $p < .001$) respectively. For GAD-7 these figures were 35% ($\chi_2(4) = 384.342$, $p < .001$) and 38% ($\chi_2(4) = 336.299$, $p < .001$). The models correctly classified 73.9% and 74.6% of cases on the PHQ-9 and 73.6% and 74.9% on the GAD-7 respectively.

Higher baseline scores were associated with increased chance of seeing RCSI, which can be explained by the criteria for reliable change. Higher sessions scores and risk sums were associated with decreased likelihood of RCSI (e.g. at session three PHQ-9 the likelihood of seeing RCSI was .808 per unit increase in session three score, and .695 per unit increase in risk sum). Higher SD was associated with increased likelihood of seeing RCSI (e.g. odds were 1.275 per unit increase in SD at session four PHQ-9). Similar odds were evidenced in the GAD-7 models.

High intensity. At high intensity, all models significantly predicted RCSI. The LRI significantly added value beyond all other variables between sessions one-six in the Complex Dynamic model. Given that therapy duration is generally longer at HIT, examples are provided from sessions six and seven, where clinicians would ideally have

a good picture of therapy response to ameliorate risk. At sessions six and seven on the PHQ-9, the Complex model explained 28% ($\chi_2(5) = 150.915$, $p < .001$) and 29% ($\chi_2(4) = 143.611$, $p < .001$) of the variance respectively. For GAD-7 these figures were 30% ($\chi_2(6) = 183.874$, $p < .001$) and 30% ($\chi_2(4) = 173.236$, $p < .001$). The models correctly classified 70.8% / 68.4% of cases on the PHQ-9 and 71.8% / 73.2% on the GAD-7.

Similar odds patterns were found as for low intensity, where higher baseline scores were associated with increased chance of seeing RCSI. Higher sessions scores were associated with reduced likelihood of RCSI (e.g. at session six PHQ-9 the likelihood of seeing RCSI was .855 per unit increase in session six score), and similarly with risk sum (.857 at session seven). Higher SD was again associated with increased likelihood of seeing RCSI (for example 1.233 per unit increase in SD at session six PHQ-9).

The LRI however differed at HIT: the odds of seeing RCSI were .751 for Moderate LRI groups compared with Low LRI, and .399 for High compared with Low LRI. The higher the LRI, the lower the chance of seeing RCSI. Similar odds patterns were found in the GAD-7 models. Histograms indicated that there was more variability in LRI groups at HIT than at LIT (appendix F).

Differences were noted between RS and SD on the different measures at HIT. RS was significant between sessions 6-12 on the PHQ-9 and 4-12 on the GAD-7. SD was significant between 3-8 on the PHQ-9 and 4-10 on the GAD-7. These variables therefore appear to be useful over a slightly broader period on the GAD-7 than the PHQ-9, and SD appears to offer more mid-range predictive value than RS.

Overall the models explained more variance around sessions four and five at LIT than at latter sessions. As discussed in the GEL literature (Barkham et al., 2006), the samples represent different aggregates of people at each session. Given the median

number of sessions was five at LIT and sample numbers were low beyond this, it is possible that the fall in variance explained is due to slower to respond samples remaining at later sessions. At HIT the variance explained does not drop, however the median number of sessions was 12. It is possible that had analyses progressed beyond this a diminishing pattern of variance explained may have been seen.

Model Comparisons (A3, RQ3, OB4).

Low intensity. The Progressive Dynamic model was considered superior at low intensity as it explained significantly more variance than more simple models but did not require the LRI. For example, by session three it was capable of explaining 32% of the variance on the PHQ-9 RCSI outcomes, compared with 5% explained by the LRI only. Although the Session Scores only model explained 30% of the variance at session three PHQ-9, the addition of the RS was significant ($\chi_2(1) = 7.583$, p=.006). Given that the RS and SD consist of relatively simple manipulations of data, these small but significant gains in predictive ability were considered useful.

High intensity. At high intensity, the Complex Dynamic model outperformed other models between sessions 1-6. For example, at session six PHQ9 the Complex model explained 28% of the variance, compared with 27% in the Progressive model (where the LRI step was significant ($\chi_2(1) = 10.663$, p=.005). This difference was even clearer when compared with the static session score and LRI only models, which explained 23% and 4% of the variance respectively. These figures and patterns were similar for GAD-7. Although the Progressive Dynamic model performed nearly as well across the dataset at both low and high intensity, the Complex model was considered superior given the inclusion of important information from the LRI at earlier stages.

Phase Two: Cross-Validating Predictive Models in the Cumbria Dataset (A2, RQ2, OB3).

Cross-validation took place using AUC, Kappa and Brier scores. At LIT, sample numbers ranged from *n*=857 at session one to *n*=152 at session seven. Sample numbers were too low at session eight to interpret (*n*=36), however data are presented for information. At HIT, sample numbers ranged from *n*=1107 at session one to *n*=125 at session 12. Data from session 12 should not therefore be interpreted as reliable. Results for AUC, Kappa and Brier analyses are presented in the graphs below including confidence intervals for Kappa and AUC. Non-significant results are indicated by striped bars in graphs and noted underneath figures. Brier scores do not have associated confidence intervals or statistics but are used as supplementary information.

Stream A: LIT PHQ-9 Figures

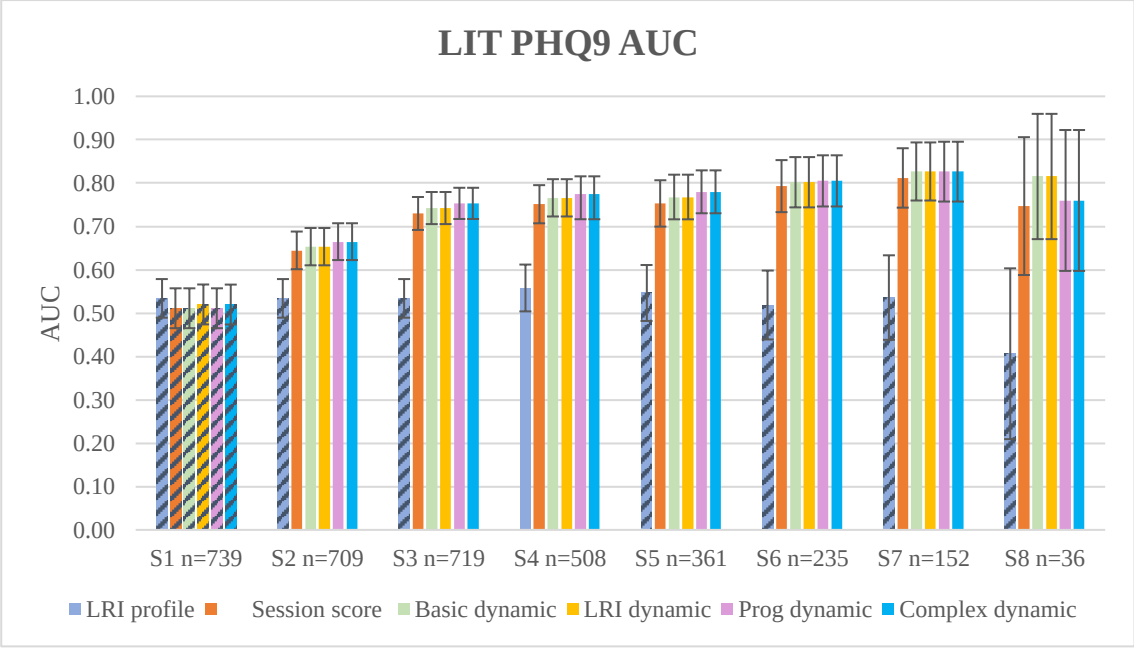


Figure 8. LIT PHQ-9 AUC.

Notes

All models are significant at the p<0.001 level apart from LRI profile only which is non-significant at all sessions apart from S4, and all other models at S1.

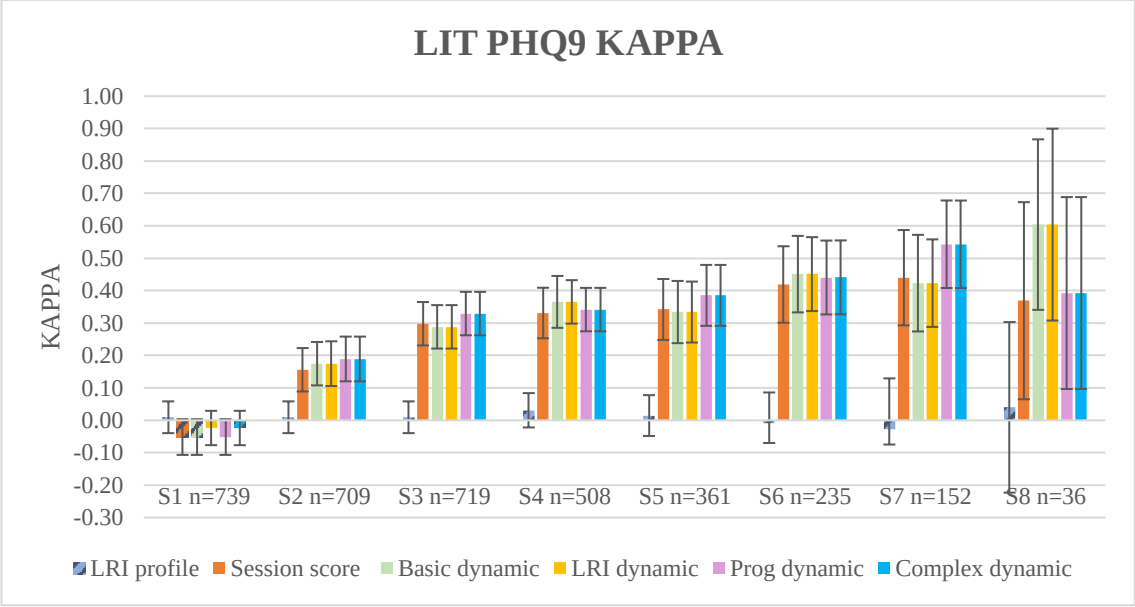


Figure 9. LIT PHQ-9 KAPPA.

Notes

All models are significant at the p<0.001 level apart from LRI profile only which is non-significant at all sessions, and all other models at S1.

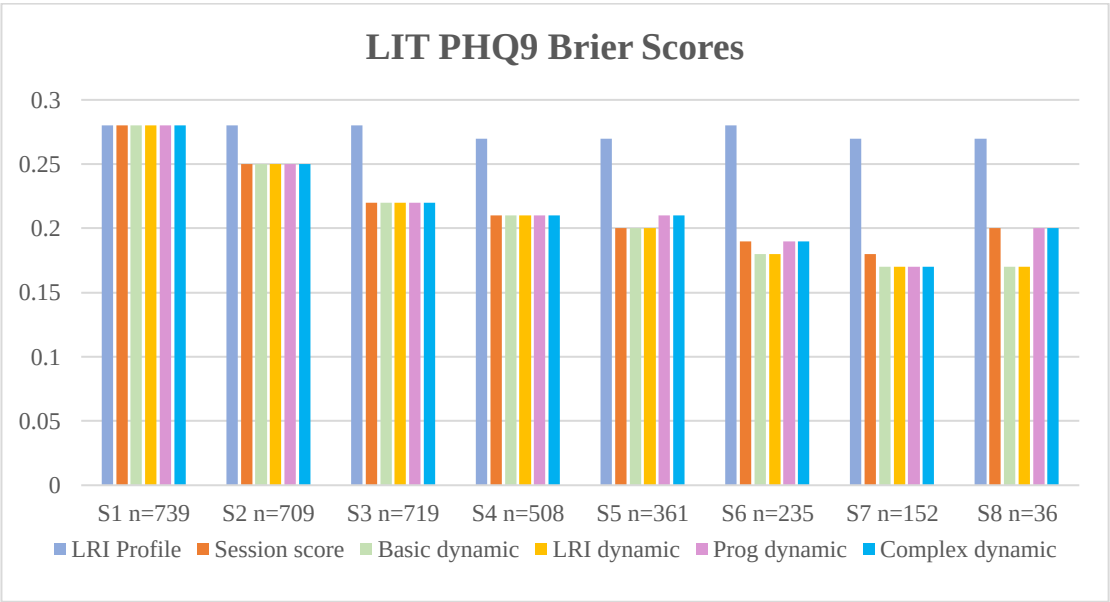


Figure 10. LIT PHQ-9 Brier scores.

Stream B: LIT GAD-7 Figures

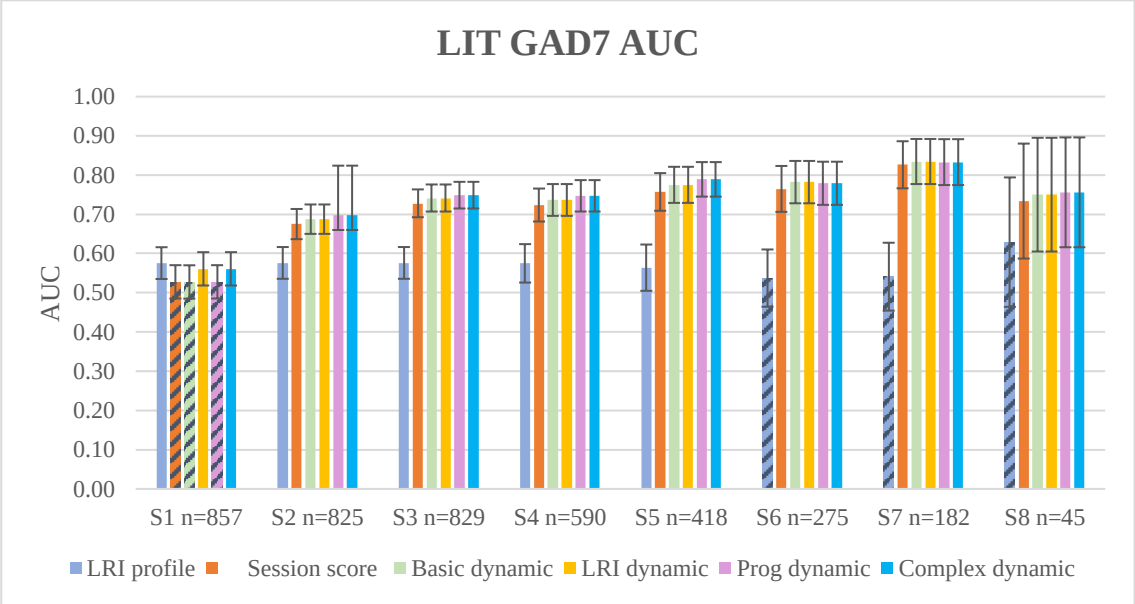


Figure 11. LIT GAD-7 AUC.

Notes.

All models are significant apart from LRI profile only which is non-significant at S6-S8, and session score, basic dynamic and progressive dynamic at S1.

All models are significant at the p<0.001 level apart from LRI dynamic and complex dynamic at S1, LRI profile at S4, and all other models at S8 (all p<.01), and LRI profile S5 (p<.05).

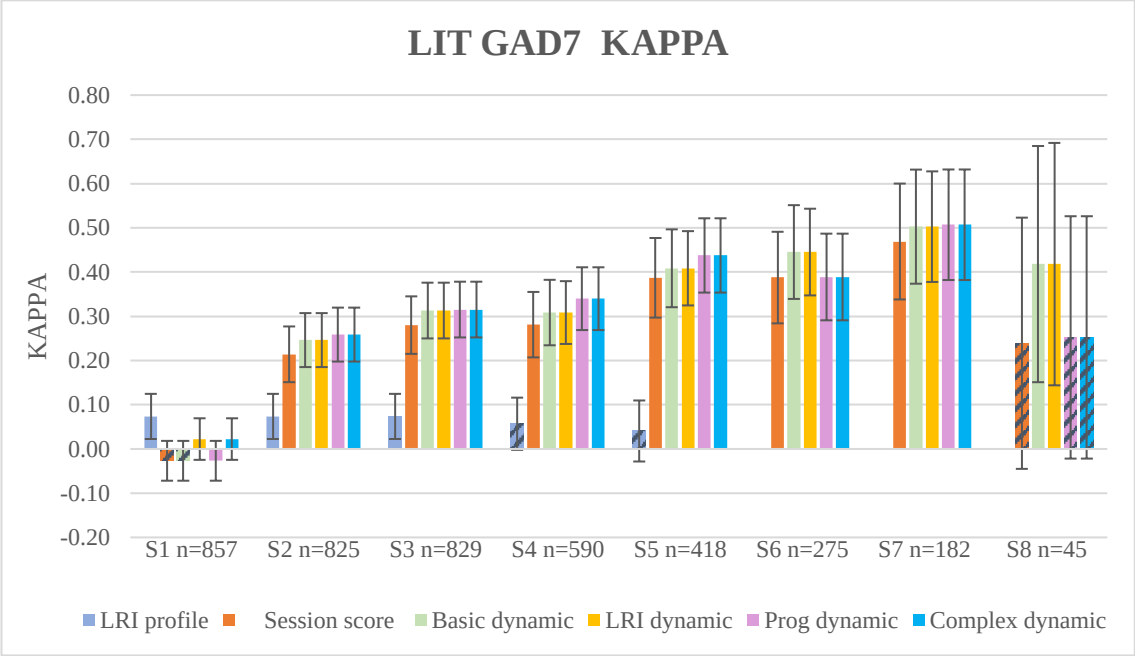


Figure 12. LIT GAD-7 KAPPA.

Notes.

All models are significant apart from LRI profile only which is non-significant at S4-S8; session score, basic dynamic, progressive dynamic and complex dynamic at S1; and session score, progressive dynamic and complex dynamic at S8.

All models are significant at the p<0.001 level apart from LRI profile S1-S3, and basic and LRI dynamic models at S8 (all p<.01).

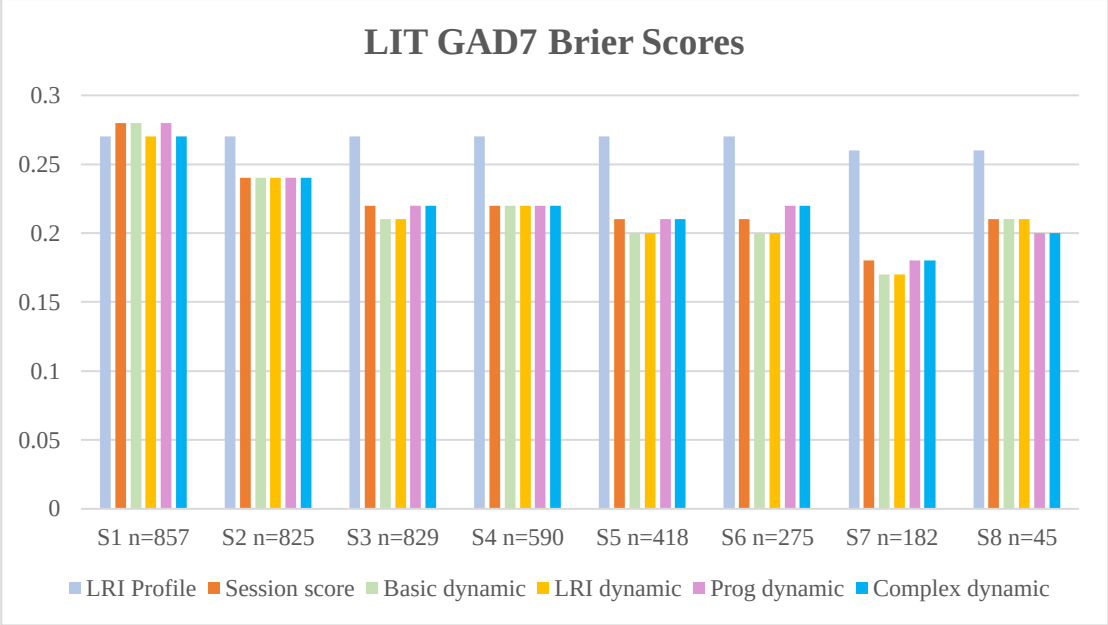


Figure 13. LIT GAD-7 Brier scores.

Stream C: HIT PHQ-9 Figures

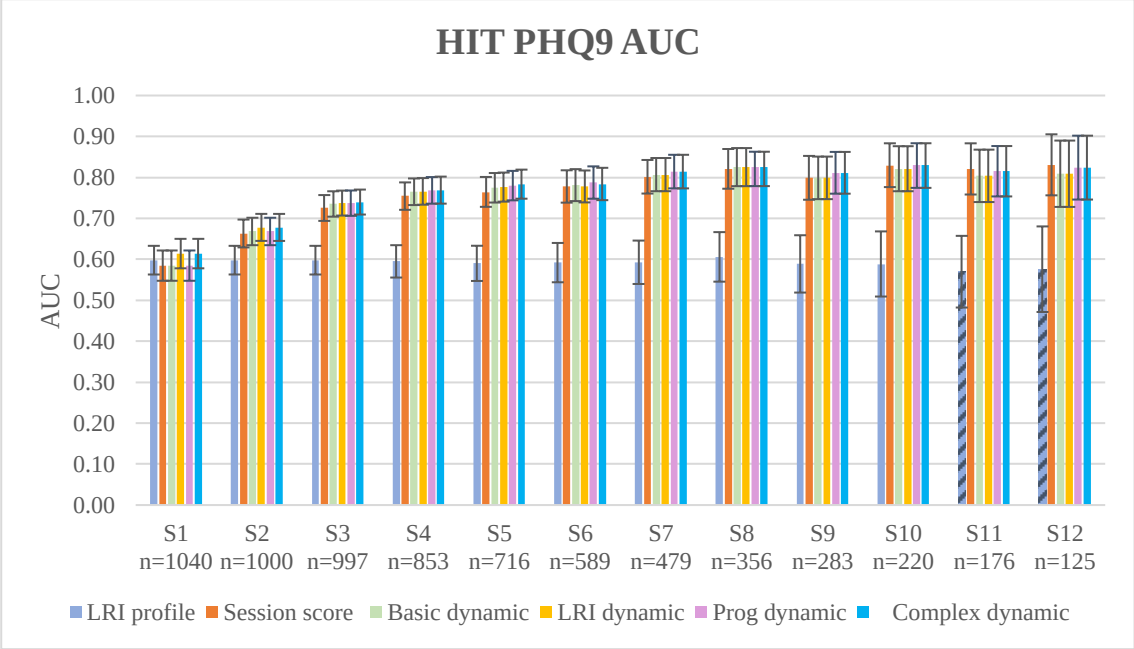


Figure 14. HIT PHQ-9 AUC.

Notes.

All models are significant apart from LRI profile only S11-S12.

All models are significant at the p<0.001 level apart from LRI profile S7-S8 (p<.01), and S9-S10 (p<.05).

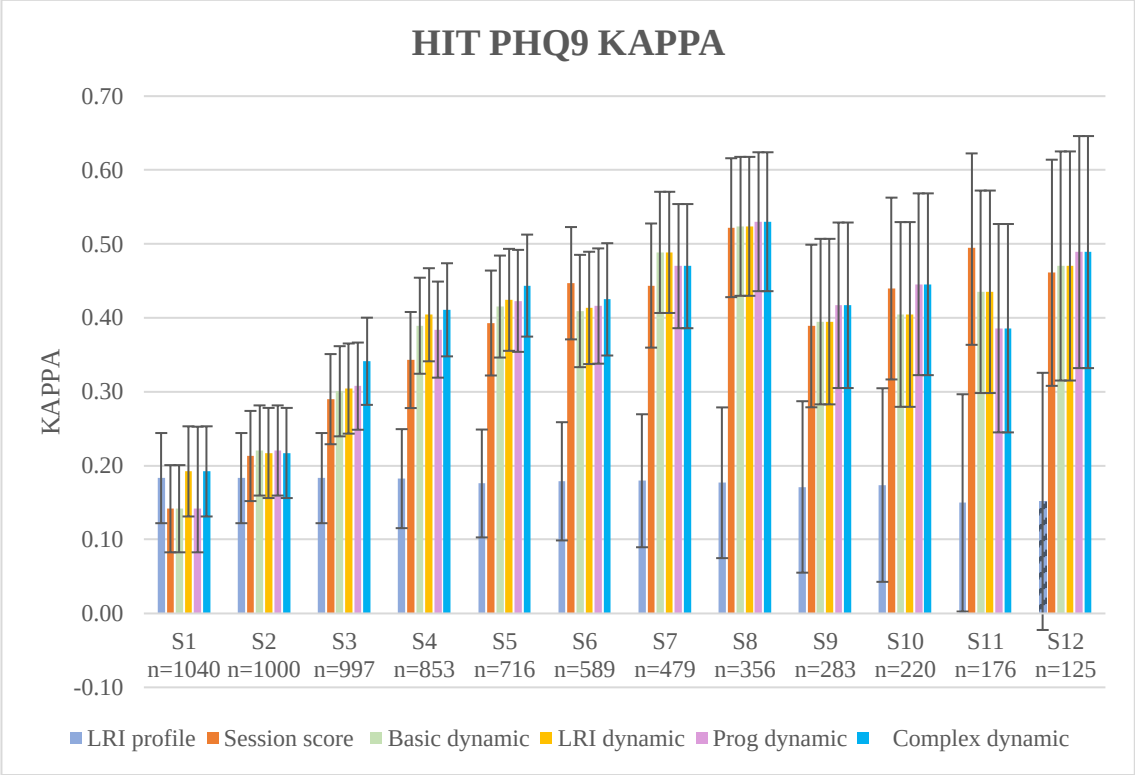


Figure 15. HIT PHQ-9 KAPPA

Notes.

All models are significant apart from LRI profile only S12.

All models are significant at the p<0.001 level apart from LRI profile S8-S10 (*p*<.01), and S11 (*p*<.05).

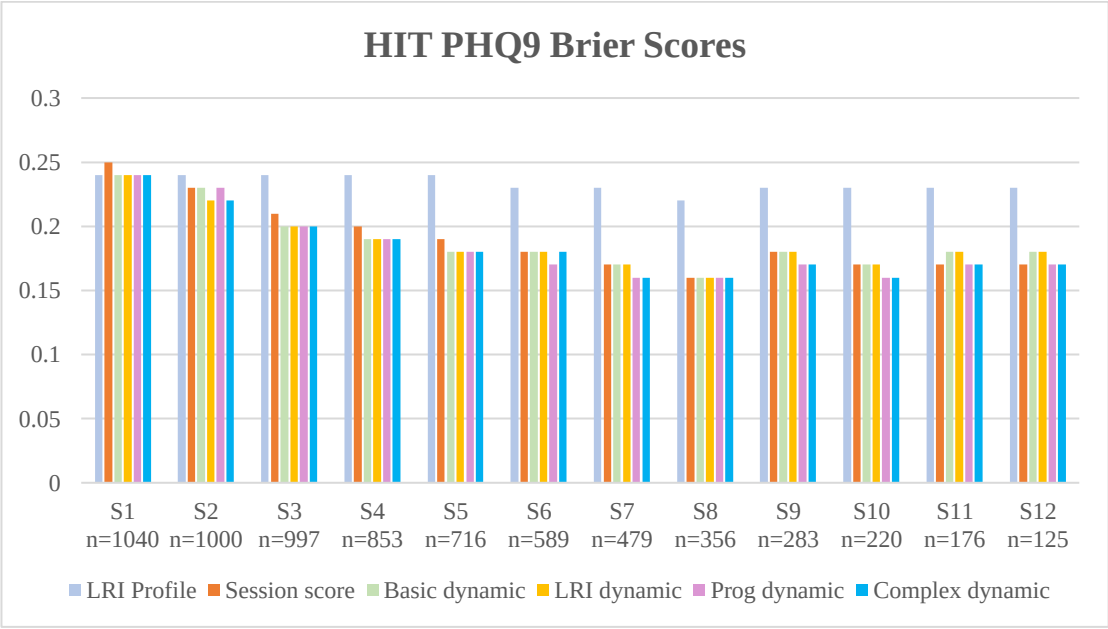


Figure 16. HIT PHQ-9 Brier scores.

Stream D: HIT GAD-7 Figures

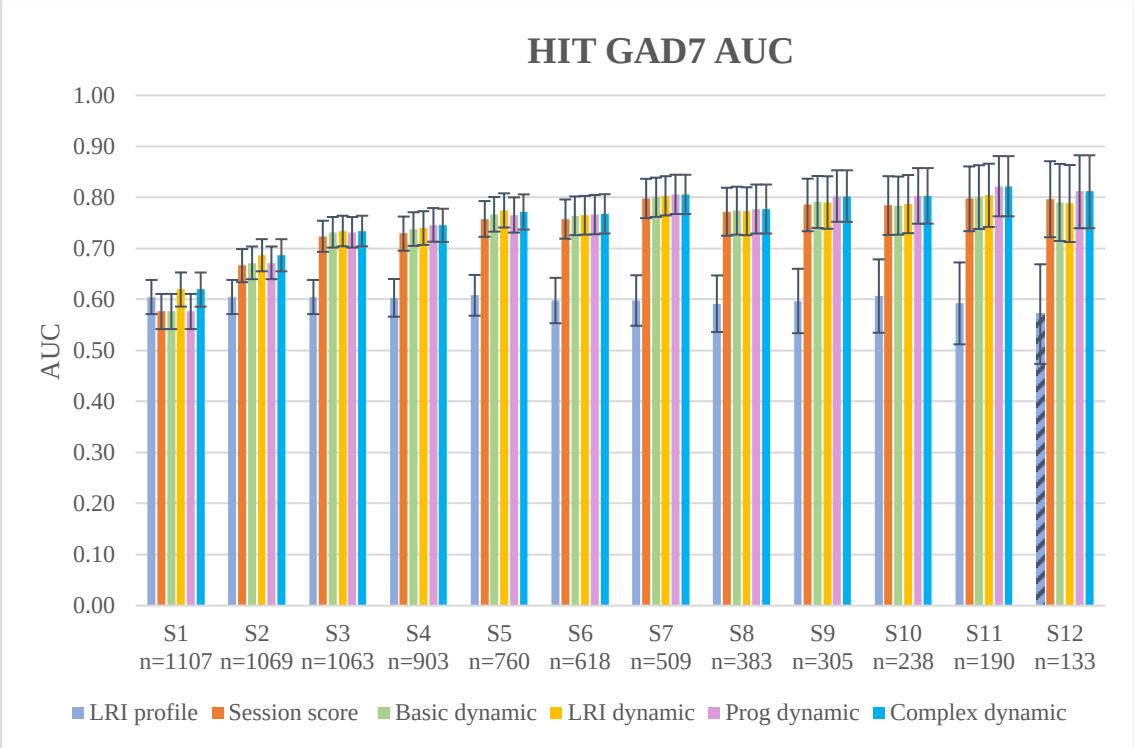


Figure 17. HIT GAD-7 AUC.

Notes.

All models are significant apart from LRI profile only S12.

All models are significant at the p<0.001 level apart from LRI profile S8-S10 (p<.01), and S11 (p<.05).

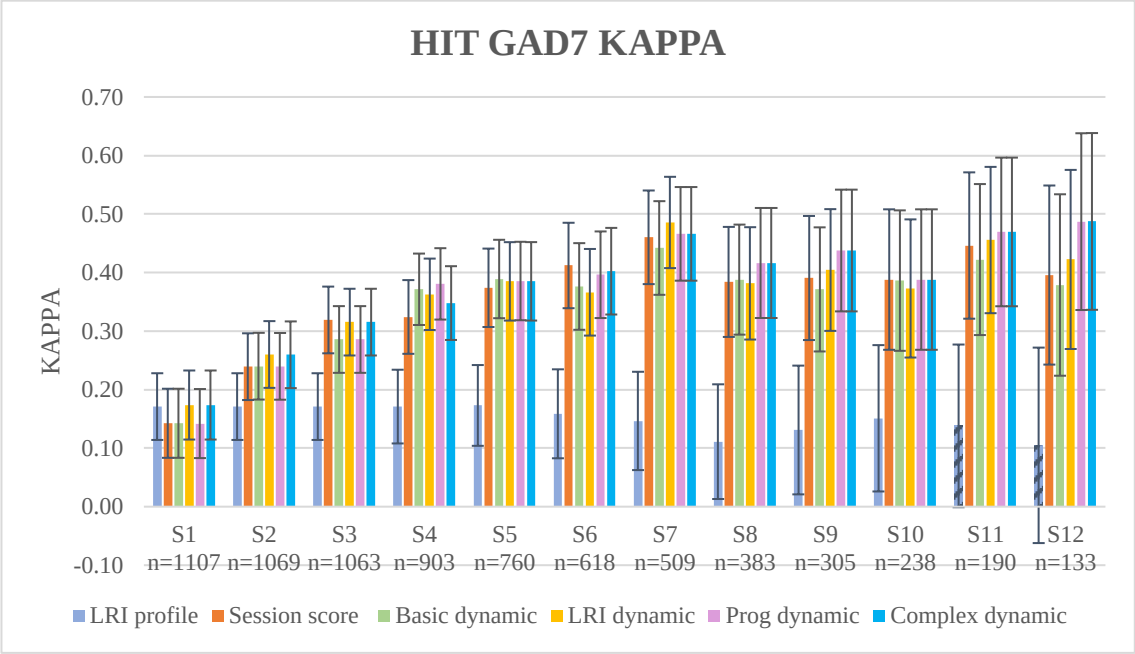


Figure 18. HIT GAD-7 KAPPA.

Notes.

All models are significant apart from LRI profile only S11-S12.

All models are significant at the p<0.001 level apart from LRI profile S7 (*p*<.01), and S8-S10 (*p*<.05).

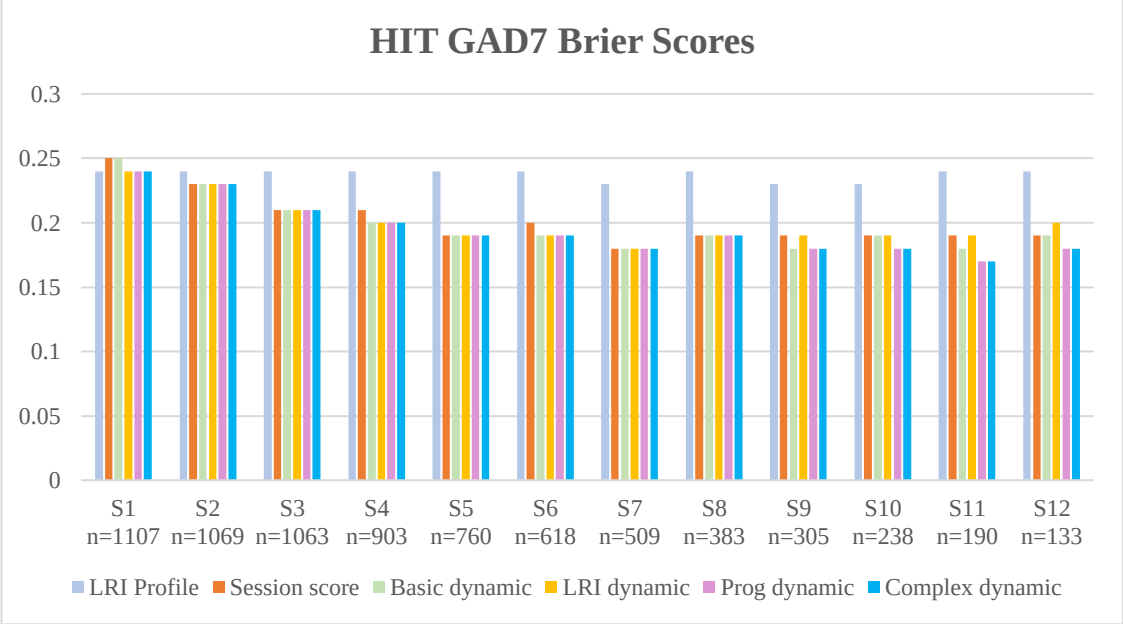


Figure 19. HIT GAD-7 Brier scores.

Summary of Cross-Validation

AUC. Hosmer & Lemeshow (2013) describe AUC values of 0.7-.08 as acceptable, where 0.5 would be chance, and above .8 excellent. AUC analyses indicated that beyond session three, all of the models, apart from LRI Profile only, had acceptable to excellent ability to correctly classify cases in the cross-validation dataset. The Progressive and Complex models were equivalent at low intensity as the LRI had not been a significant predictor beyond session one. For example, at low intensity, AUC of .753 /.775 was found at sessions three/four on PHQ-9 outcomes respectively for both these models. The figures for GAD-7 were AUC .749 / .747.

At high intensity, the complex model outperformed other models up to session six, before the Progressive model became superior (showing an AUC of .788 / .814 at session six/seven on the PHQ-9). A similar pattern was observed on the GAD-7. All of the models' confidence intervals overlapped apart from the baseline LRI profile model, however. This suggests that both the static session score and more complex models perform well in the cross-validation dataset.

Kappa. Kappa results suggested a fair to moderate ability to correctly classify cases apart from LRI profile only, which performed worse (Altman, 1999). The Progressive and Complex models tended to see higher Kappa figures and were equivalent at LIT. For example, by session three on LIT PHQ-9, a Kappa of 0.329 (fair) was found for the Progressive/Complex models, where chance agreement would be zero and negative values would be disagreement ($\kappa = .329$, $p < .001$). Similar results were found for GAD-7. For HIT PHQ-9, a Kappa of 0.425 (moderate) was found at session six ($\kappa = .416$, $p < .001$) in the complex model, which slightly outperformed all others. Again, similar figures were seen in the GAD-7.

Generally, greater variability was seen in the Kappa scores. As with the AUC, confidence intervals overlapped across models indicating that the more complex dynamic models were not significantly better than basic ones beyond the LRI only. Feinstein & Cicchetti (1990) discussed a possible limitation of Kappa ('the paradox') however, due to its dependence on marginal distributions, where Kappa scores cannot be compared unless the marginal distributions are the same (base rates of RCSI in the samples). This may explain the findings here and makes Kappa more difficult to interpret and compare.

Brier scores. There are no established criteria for interpreting Brier scores. However 1 is considered to be perfectly inaccurate and 0 perfectly accurate. At LIT PHQ-9, Brier scores at session three/four were 0.22 and 0.21 respectively for all models apart from LRI profile only (which was higher at 0.28). Similar scores were found on the GAD-7. Note that at sessions five and six, the basic dynamic and LRI dynamic models marginally outperformed the progressive and complex models.

At HIT PHQ-9, a Brier score of 0.17 was found at session six for the progressive model and 0.18 on all others apart from LRI profile. The progressive model also had the lowest score of 0.16 at session seven. The figures for GAD-7 at sessions six-seven were slightly higher at 0.19 and 0.18. Overall there were only slight differences between all models apart from the LRI profile only, which performed worse (it being more distal). These scores would suggest that error was relatively low in all of the models, with the highest error being in more distal (early) predictions at LIT, and least error being at sessions seven or eight at HIT.

Discussion

Model Development

This study examined whether client profile and routine psychotherapy outcomes measures could be integrated into dynamic prediction systems. Findings indicated that it was possible to build dynamic models that significantly improved on static methods in the development phases. The LRI did not add value beyond the other variables at LIT. However at sessions one-six at HIT, it did. This suggests that additional elements captured within the LRI are more important predictors for people accessing HIT than LIT. As discussed, there was greater variability in LRI scores at HIT, facilitating better discrimination between cases. It is also understandable that the LRI would add more value in earlier stages, where over the course of therapy early profile information becomes less relevant.

The RS and SD are cumulative predictors, therefore it is understandable that they might increase in predictive value over time. However whereas the RS added value up to session 12 on both measures, SD offered more mid-range benefit. This could be explained by the fact that later samples may include those making less change (less variability from their own mean).

Cross-Validation

The second aim was to examine whether such models could be cross-validated in a new sample. All analyses suggested a moderate to excellent ability to predict outcomes in a new sample, apart from the LRI only model. The Progressive and Complex models saw higher figures than the other models; however confidence intervals overlapped suggesting differences were not significant. Taken together, these cross-validation analyses provided some confidence that the dynamic models could generalise to a new sample.

Is Complexity Better?

The final aim was to consider whether complex models outperformed static or more basic dynamic models. At LIT, the Progressive Dynamic was arguably the preferred model as the LRI did not add value beyond session one. However at HIT the Complex Dynamic model outperformed other models at earlier sessions. In the cross-validation dataset the progressive and complex models tended to see higher figures on the AUC and Kappa between sessions one-six, however differences were small and confidence intervals overlapped.

If the preference is for parsimony, the Progressive model would be considered pragmatic as it only requires progress measures which are already routinely collected and easily computed through data manipulations. Nonetheless, where services have been able to implement the LRI unproblematically, there is the possibility for increased value in identifying risk in earlier sessions. Although the differences between models were not statistically significant in the cross-validation sample, the figures tended nonetheless to be higher for the complex model at earlier sessions. On the ground it is possible that a number of people could still be identified and supported who may be missed without this information. The answer therefore depends on how difficult it might be for individual services to adopt and integrate the LRI into IT systems.

Limitations

Several limitations are noted, which relate to methods of categorising cases, sample demographics, pragmatic choices about predictors, and considerations for interpreting outputs. As discussed, cases at LIT and HIT were treated as separate incidents of therapy, however some people may have accessed low intensity and stepped up to high intensity and this could have been examined as a predictor. However sample numbers would have been low at each session and previous research in IAPT

indicates that preceding information from low intensity treatment does not predict outcomes at high intensity (Delgadillo et al., 2017).

There might be other factors that predict whether someone does well in therapy that were not included in the models, such as therapist or intervention effects. However, models including less fixed aspects such as these would also be less generalisable to other services or future samples, where therapists and intervention details vary and change. Some presenting problems may have poorer prognoses than others, however sample sizes are often smaller for these factors and results may be biased as a result. People may also have overlapping problems or problems that change during the course of therapy, and pre-defined presenting problems are not always helpful in practice.

A further issue is that although the models have cross-validated in a further sample, there may be characteristics that both the development and cross-validation samples have in common. For example, they are both Northern UK IAPT services with a limited multi-ethnic demographic. However they both also differed in clear ways, where the Leeds dataset had lower RCSI rates than Cumbria, particularly at LIT, and made more use of group interventions. Cumbria also appeared to match high and low LRI groups to high and low intensity more consistently than the Leeds dataset, which could relate to differences in how services currently stratify care and would be important to consider in future research.

Missing data in the study were treated as missing at random, and only contacts actually attended were included in the analyses. This is understandable given the need for outcome measures inherent in patient-focused research which is dependent on 'completers': if we do not have outcome measures, we cannot measure an outcome. However missing data may not be random and may provide important information about outcomes (Gottfredson, Bauer, Baldwin, & Okiishi, 2014). Imputation was not

considered necessary with such a large dataset, and arguably addresses a different issue. Thus current modelling techniques in general remain limited in predicting future trajectories for those who are most likely to drop-out.

A further difficulty of application for the models is the concept of differential patterns of change. The models provide personalised risk predictions based on comparisons with others having similar dynamic profiles, and can capture elements such as sudden gains (Delgadillo et al., 2014; Lutz et al., 2012). However research using GMM techniques has illustrated that different sub-classes can be identified who follow atypical trajectories (Rubel et al., 2015), including 'worse before better' (Owen et al., 2015). In this last example, a dip would not necessarily mean someone was "not on track". If this was a common pattern the models would accommodate it, however if this occurred in a smaller subset of people, the average response would shape the predicted output leading to possible prediction error.

This is difficult to overcome, because although we can identify sub-classes of people retrospectively, we currently remain dependent on profiling techniques at outset and cannot reliably predict heterogeneity in shapes of change. This relates to the issue of how the models could be used in clinical practice and is discussed further below. There are also further limitations beyond the scope of this paper, such as critiques of the measurement of outcomes more broadly (see Ogles, 2013). However the models here were oriented to supporting a system that already exists, and measurements are not intended to be used in isolation but to prompt clinical conversations as discussed below.

Strengths

This study is the first to combine baseline profile information from the LRI with incoming weekly progress scores into a dynamic prediction system. As discussed in the introduction, current systems tend to rely on visual inspection of progress scores, or

ETR curves based on group norms; however the GEL literature has shown that not everyone responds according to dose-response curves (see Castonguay, Barkham, Lutz, & McAleavey, 2013). This system overlaps to some extent, where the models compare individual data against a cohort; however this system is also different. It not only combines LRI profile information with incoming scores, but it includes learning from previous sessions in the form of sample level risk sums and individual standard deviations. Further, the system does not suppose a fixed pattern of response but provides individualised and recalculated probabilities of seeing improvement as new information comes in. This therefore offers a different way of alerting clinicians to who is "not on track" through not only indicating a risk, but including quantification of an individual's chance of seeing RCSI. The models were able to show good ability to predict outcomes in a new dataset, which is particularly encouraging given some of the heterogeneity in characteristics and outcomes observed between the two services, which is typical in practice (Clark et al., 2018).

The models were discussed as potentially being limited by not including all possible predictors. However the addition of too many predictors into models can also lead to the problem of 'overfitting', where data do not generalise well to other samples and settings (Field, 2018). Given a preference for parsimony, the simplicity of using profile information in combination with incoming progress scores make these models more likely to be useful in a wide range of settings, easier to translate into IT systems, and outputs more readily understood by clinicians.

Theoretical Implications and Future Research

The generalisability of these models is important for them to be useful in routine practice, and there are plans for further development of these algorithms using machine learning techniques (such as LASSO or Elastic Net Regularisation) to minimise

overfitting (see Cohen, Kim, Van, Dekker, & Driessen, 2019; Delgadillo et al., 2017) in a wider range of IAPT services. Further, whilst there is a growing body of evidence linking feedback methods to improved outcomes, it would be important to examine whether or not providing a prediction system such as this actually benefits clients and services in practice.

Future research would also be important to assess differential patterns of change within predictive modelling to assist with interpretation of model outputs. For example, if it was possible to identify sub-classes of responders that could generalise to new samples clinicians would be aided in interpreting outputs. Research on the GEL models suggests that faster responders or certain problem types may see a more curvilinear response to treatment, whilst slower responders may appear more linear (Kivlighan et al., 2018; Owen et al., 2016). However this is not currently well understood and is contradicted in some cases.

It would also be helpful to examine the impact of missing data on outcomes prediction systems and how this shapes current understanding of patient trajectories. Current understanding is based on cohorts with completed outcome measures, however to be better able to predict outcomes in future clients more needs to be known about non-completers (Gottfredson et al., 2014).

Clinical Implications

These models have been developed in collaboration with IAPT Leeds and Cumbria, with the intention of considering future clinical application. An important issue therefore is how they would be communicated and used. For example, they are intended as a form of ‘SatNav’ system, to alert clinicians to who is most at risk of not seeing improvement, based on the best information available and not limited by assumptions of decelerating change trajectories. It is possible that the clinicians could

see a rank percentage order of who is most at risk of poor outcomes, which could include graphs charting individual progress to understand the trend. Future steps could therefore include the development of a flowchart to work alongside the prediction system to support clinicians.

Ultimately they are intended to prompt thinking and conversations between therapist-supervisor and client-therapist to consider why a pattern might be emerging. For example, is it a 'worse before better' pattern because someone is engaging well but working through something difficult (Owen et al., 2015), or is it a deterioration that may predict poor outcomes? (Lutz et al. (2012). These would be good questions to discuss with clients and use in supervision, given the overall objectives of increasing client wellbeing and preventing re-referrals.

This heterogeneity of response reflects the wider debate on whether humans can be studied nomothetically or idiographically, whether universal laws can be identified (positivism) or whether people are entirely unpredictable (indeterminism, Cziko, 1989). It seems likely that we are both, we can be predicted and we share features in common with one another, however within certain parameters we are also unpredictable or exhibit differences. There is an argument that instead of trying to see unexplained variance as undesirable, it indicates a human possibility for unpredictable and random behaviour, where unpredictable or protean behaviour also implies the possibility for change.

Conclusions

This study provided an illustration of how baseline patient profile information can be integrated with incoming progress measures to provide dynamic progress feedback systems. These systems were capable of providing personalised prognoses of outcome and percentage likelihoods of achieving this, learning and updating as new

information was accumulated. The models showed good evidence of generalising to another sample, however further cross-validation and the examination of impact on outcomes in practice is intended. Overall, the complex model tended to outperform more basic models, particularly in earlier sessions at HIT, however confidence intervals overlapped in the cross-validation sample. Preference for complex over more basic models may therefore depend on service capabilities.

References

Altman, D. G. (1999). *Practical statistics for medical research*. New York, NY: Chapman & Hall/CRC Press.

Baldwin, S. A., Berkeljon, A., Atkins, D. C., Olsen, J. A., & Nielsen, S. L. (2009). Rates of change in naturalistic psychotherapy: Contrasting dose-effect and good-enough level models of change. *Journal of Consulting and Clinical Psychology, 77*, 203–211. https://doi.org/10.1037/a0015235

Barkham, M., Connell, J., Stiles, W., Miles, J., Margison, F., Evans, C., . . . & La Greca, A. M. (2006). Dose-effect relations and responsive regulation of treatment duration: The good enough level. *Journal of Consulting and Clinical Psychology, 74*, 160-167. DOI:10.1037/0022-006X.74.1.160

Barkham, M., Rees, A., Stiles, W., Shapiro, D., Hardy, G., & Reynolds, S. (1996). Dose-effect relations in time-limited psychotherapy for depression. *Journal of Consulting and Clinical Psychology, 64*, 927-935. http://dx.doi.org/10.1037/0022-006X.64.5.927

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*, 1-3. Retrieved from: ftp://ftp.library.noaa.gov/docs.lib/htdocs/rescue/mwr/078/mwr-078-01-0001.pdf

Castonguay, L., Barkham, M., Lutz, W., & McAleavey, A. (2013). Practice-oriented research: Approaches and applications. In M. Lambert (Ed.). *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th Ed., pp. 85-133). New Jersey, NJ: John Wiley & Sons, Inc.

Clark, D. M. (2011). Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: The IAPT experience. *International Review of Psychiatry, 23*, 318–327. https://doi.org/10.3109/09540261.2011.606803

Clark, D. M., Canvin, L., Green, J., Layard, R., Pilling, S., Janecka, M. (2018). Transparency about the outcomes of mental health services (IAPT approach): an analysis of public data. *The Lancet, 391*, 679-686. https://doi.org/10.1016/S0140-6736(17)32133-5

Cohen, Z. D., Kim, T.T., Van, H. L., Dekker, J. J. M., & Driessen, E. (2019). A demonstration of a multi-method variable selection approach for treatment selection: Recommending cognitive–behavioral versus psychodynamic therapy for mild to moderate adult depression. *Psychotherapy Research.* https://doi.org/10.1080/10503307.2018.1563312

Cziko, G. A. (1989). Unpredictability and indeterminism in human behaviour: Arguments and implications for educational research. *Educational Researcher, 18*, 17-25. Retrieved from: http://www.jstor.org/stable/1174887

Delgadillo, J., de Jong, K., Lucock, M., Lutz, W., Rubel, J., Gilbody, S., … & Mcmillan, D. (2018). Feedback-informed treatment versus usual psychological treatment for depression and anxiety: A multisite, open-label, cluster randomised controlled trial. *The Lancet Psychiatry, 5*, 564–72. https://doi.org/10.1016/S2215-0366(18)30162-7

Delgadillo, J., Huey, D., Bennett, H., & McMillan, D. (2017). Case complexity as a guide for psychological treatment selection. *Journal of Consulting and Clinical Psychology, 85*, 835-853. http://dx.doi.org/10.1037/ccp0000231

Delgadillo, J., McMillan, D., Lucock, M., Leach, C., Ali, S., & Gilbody, S. (2014). Early changes, attrition, and dose-response in low intensity psychological interventions. *British Journal of Clinical Psychology, 53,* 114–130. https://doi.org/10.1111/bjc.12031

Delgadillo, J., Moreea, O., & Lutz, W. (2016). Different people respond differently to therapy: A demonstration using patient profiling and risk stratification. *Behaviour Research and Therapy, 79,* 15-22. http://dx.doi.org/10.1016/j.brat.2016.02.003

Evans, C., Margison, F., & Barkham, M. (1998). The contribution of reliable and clinically significant change methods to evidence-based mental health. *Evidence Based Mental Health, 1,* 70-72. http://dx.doi.org/10.1136/ebmh.1.3.70

Feinstein. A.R, & Cicchetti. D.V. (1990). High agreement but low kappa: I. The Problems of Two Paradoxes. *Journal of Clinical Epidemiology, 43,* 543-548. https://doi.org/10.1016/0895-4356(90)90158-L

Field, A. (2018). Discovering statistics using IBM SPSS statistics (5th Ed.). London, UK: SAGE Publications Ltd.

Gottfredson, N. C., Bauer, D. J., Baldwin, S. A., & Okiishi, J. C. (2014). Using a shared parameter mixture model to estimate change during treatment when termination is related to recovery speed. *Journal of Consulting and Clinical Psychology, 82*, 813-827. http://dx.doi.org/10.1037/a0034831

Hannan, C., Lambert, M.J., Harmon, C., Nielson, S.L., Smart, D.W., Shimokawa, K., & Sutton, S.W. (2005). A lab test and algorithms for identifying clients at risk or treatment failure. *Journal of Clinical Psychology, 61*, 155-163. https://doi.org/10.1002/jclp.20108

Hosmer, D.W., & Lemeshow, S. (2013). *Applied logistic regression* (3rd Ed.). New Jersey: John Wiley & Sons, Inc.

Howard, K.I., Kopta, S.M., Krause, M.S., & Orlinksy, D.E. (1986). The dose-effect relationship in psychotherapy. *American Psychologist, 41*, 159-164. http://dx.doi.org/10.1037/0003-066X.41.2.159

Hsieh, F. Y. (1989). Sample size tables for logistic regression. *Statistics in Medicine, 8,* 795-802. https://doi.org/10.1002/sim.4780080704

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12-19. http://dx.doi.org/10.1037/0022-006X.59.1.12

Kivlighan, D. M., Lin, Y. J., Egan, K. P., Pickett, T., & Goldberg, S. B. (2018). A further investigation of the good-enough level model across outcome domains and termination status. *Psychotherapy, (advance online publication)*. http://dx.doi.org/10.1037/pst0000197

Kroenke, K., Spitzer, R.L., & Williams, J.B.W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine, 16*, 606-613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Lambert, M., Whipple, J., Kleinstäuber, M., Hilsenroth, Mark J., & Norcross, J. C. (2018). Collecting and Delivering Progress Feedback: A Meta-Analysis of Routine Outcome Monitoring. *Psychotherapy, 55*, 520-537. http://dx.doi.org/10.1037/pst0000167

Lutz, W. (2002). Patient-focused psychotherapy research and individual treatment progress as scientific groundwork for an empirically based clinical practice. *Psychotherapy Research, 12,* 251-272. https://doi.org/10.1080/713664389

Lutz, W., Leach, C., Barkham, M., Lucock, M., Stiles, W. B., Evans, C., . . . & Iveson, S. (2005). Predicting change for individual psychotherapy clients on the basis of their nearest neighbors. *Journal of Consulting and Clinical Psychology, 73*(5), 904-913. http://dx.doi.org/10.1037/0022-006X.73.5.904

Lutz, W., Ehrlich, T., Rubel, J., Hallwachs, N., Röttger, M., Jorasz, C., . . . & Tschitsaz-Stucki, A. (2012). The ups and downs of psychotherapy: Sudden gains and sudden losses identified with session reports. *Psychotherapy Research, 23,* 1-11. https://doi-org.sheffield.idm.oclc.org/10.1080/10503307.2012.693837

Lutz, W., Stultz, N., Martinovich, Z., Leon, S., & Saunders, S. (2009). Methodological background of decision rules and feedback tools for outcomes management in psychotherapy. *Psychotherapy Research, 19,* 502-510. https://doi.org/10.1080/10503300802688486

National Institute for Health and Care Excellence (2011). *Common mental health disorders: identification and pathways to care.* Clinical guideline [CG123]. Retrieved from: http://guidance.nice.org.uk/CG123/NICEGuidance/pdf/English

NHS Digital (2018). Psychological therapies, annual report on the use of IAPT services – England, 2017-18. Retrieved from: https://digital.nhs.uk/data-and-information/publications/statistical/psychological-therapies-annual-reports-on-the-use-of-iapt-services/annual-report-2017---18

Ogles, B. M. (2013). Measuring change in psychotherapy research. In M. Lambert (Ed.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change*, (6th Ed., pp. 134-166). New Jersey, NJ: John Wiley & Sons, Inc.

Owen, J., Adelson, J., Budge, S., Wampold, B., Kopta, M., Minami, T., & Miller, S. (2015). Trajectories of change in psychotherapy. *Journal of Clinical Psychology, 71*, 817-827. https://doi.org/10.1002/jclp.22191

Owen, J., Adelson, J., Budge, S., Kopta, S. M., & Reese, R. J. (2016). Good-enough level and dose-effect models: Variation among outcomes and therapists. *Psychotherapy Research*, *26*, 22–30. https://doi.org/10.1080/10503307.2014.966346

Redelmeier, D. A., Block, D. A., & Hickam, D. H. (1991). Assessing predictive accuracy: How to compare Brier scores. *Journal of Clinical Epidemiology, 44*, 1141-1146. https://doi-org.sheffield.idm.oclc.org/10.1016/0895-4356(91)90146-Z

Richards, D. A., & Borgin, G. (2011). Implementation of psychological therapies for anxiety and depression in routine practice: Two year prospective cohort study. *Journal of Affective Disorders, 133*, 51-60. https://doi.org/10.1016/j.jad.2011.03.024

Rubel, J., Lutz, W., & Schulte, D. (2015). Patterns of change in different phases of outpatient psychotherapy: A stage-sequential pattern analysis of change in session reports. *Clinical Psychology and Psychotherapy, 22*, 1-14. http://dx.doi.org/10.1002/cpp.1868

Rufibach, K. (2010). Use of Brier score to assess binary predictions. *Journal of Clinical Epidemiology, 63*, 938-942. https://doi.org/10.1016/j.jclinepi.2009.11.009

Saunders, R., Cape, J., Fearon, P., & Pilling, S. (2016). Predicting treatment outcome in psychological treatment services by identifying latent profiles of patients.

Journal of Affective Disorders, 197, 107-115. https://doi.org/10.1016/j.jad.2016.03.011

Saxon, D., & Barkham, M. (2012) Patterns of therapist variability: Therapist effects and the contribution of patient severity and risk. *Journal of Consulting and Clinical Psycholology 80*, 535-546. https://doi.org/10.1037/a0028898

Spitzer, R.L., Kroenke, K., Williams, J.B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine, 166,* 1092-1097. https://doi.org/10.1001/archinte.166.10.1092

Tetlock, P., & Gardner, D. (2015). *Super-forecasting. The art and science of prediction*. London, UK: Random House Books.

Watson, P.F., & Petrie, A. (2010). Method agreement analysis: A review of correct methodology. *Theriogenology, 73,* 1167-1179. https://doi.org/10.1016/j.theriogenology.2010.01.003

Appendix A

Results Tables

Differences between included and excluded samples.

Q-Q plots indicated data were sufficiently normally distributed. Baselines were all higher in included samples. Age was significantly lower in the Leeds included sample at Step 3. Age was significantly higher in Cumbria included sample at Step 2. Males were significantly less likely to be in the included sample in Leeds Step 3.

	Leeds	Cumbria
Step 2 Baseline PHQ-9	$t(2025)=-6.936$, $p<.001$ Mean 16.22 versus 14.48	$t(1822)=-15.178$, $p<.001$ Mean 15.44 versus 11.81
Step 2 Baseline GAD-7	$t(2170)=-4.318$, $p<.001$ 14.29 versus 13.39	$t(1907)=-12.599$, $p<.001$ 13.97 versus 11.31
Step 3 Baseline PHQ-9	$t(1110)=-5.177$, $p<.001$ 16.78 versus 14.50	$t(1384)=-9.064$, $p<.001$ 17.22 versus 14.81
Step 3 Baseline GAD-7	$t(1225)=-3.874$, $p<.001$ 14.82 versus 13.38	$t(1353)=-8.411$, $p<.001$ 15.24 versus 13.30
Age Step 2	$t(4735)=-.146$, $p=.884$ 36.90 versus 36.96	$t(4829)=-2.233$, $p=.026$ 42.08 versus 40.78
Age Step 3	t (4735)=2.200, p=0.022 36.06 versus 37.14	$t(2098)=1.069$, $p=.285$ 40.61 versus 41.16
Gender Step 2	$X_2(1)=.579$, $p=.447$	$X_2(2)=1.564$, $p=.457$
Gender Step 3	$X_2(1)=6.876$, $p=.009$ 30.5% males included (versus 34.9%) 69.5% females included (versus 65.1%)	$X_2(2)=4.282$, $p=.118$

Results tables for statistical models

All significant at $p<.001$ unless marked *<.01 **<.05 or grey for non-significant.

LIT PHQ-9 Models

LIT PHQ-9 LRI Profile Model							
Model development				Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=1197	0.049	44.130	60.3	0.535	0.009	0.28	S1 *n*=739
S2 *n*=1117	0.049	44.130	60.3	0.535	0.009	0.28	S2 *n*=709
S3 *n*=1138	0.049	44.130	60.3	0.535	0.009	0.28	S3 *n*=719
S4 *n*=921	0.037	26.866	58.2	0.558	0.031	0.27	S4 *n*=508
S5 *n*=665	0.035	18.211	57.8	0.547	0.015	0.27	S5 *n*=361
S6 *n*=345	0.038	*10.103	59.4	0.518	-0.008	0.28	S6 *n*=235
S7 *n*=187	0.057	**8.546	59.6	0.535	-0.027	0.27	S7 *n*=152
S8 *n*=97	0.052	4.085	56.9	0.406	0.040	0.27	S8 *n*=36

LIT PHQ-9 Session Scores only Model							
Model development				Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=1197	0.056	50.386	58.1	0.511	-0.052	0.28	S1 *n*=739
S2 *n*=1117	0.178	158.236	65.8	0.644	0.156	0.25	S2 *n*=709
S3 *n*=1138	0.302	289.049	72.3	0.730	0.298	0.22	S3 *n*=719
S4 *n*=921	0.359	287.373	72.7	0.751	0.331	0.21	S4 *n*=508
S5 *n*=665	0.348	200.208	73.2	0.752	0.342	0.2	S5 *n*=361
S6 *n*=345	0.284	81.989	71.9	0.793	0.419	0.19	S6 *n*=235
S7 *n*=187	0.211	32.114	67.9	0.811	0.440	0.18	S7 *n*=152
S8 *n*=97	0.188	14.761	68	0.747	0.369	0.2	S8 *n*=36

LIT PHQ-9 Basic Dynamic Model (BL & SS at each session)							
	Model development			Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=1197	0.056	50.386	58.1	0.511	-0.052	0.28	S1 *n*=739
S2 *n*=1117	0.18	159.936	66.5	0.653	0.175	0.25	S2 *n*=709
S3 *n*=1138	0.308	295.679	72.8	0.742	0.288	0.22	S3 *n*=719
S4 *n*=921	0.371	298.706	74.5	0.766	0.365	0.21	S4 *n*=508
S5 *n*=665	0.359	207.445	74.3	0.767	0.334	0.2	S5 *n*=361
S6 *n*=345	0.3	87.498	74.8	0.802	0.451	0.18	S6 *n*=235
S7 *n*=187	0.235	36.031	69	0.827	0.423	0.17	S7 *n*=152
S8 *n*=97	0.243	19.523	68	0.816	0.604	0.17	S8 *n*=36

LIT PHQ-9 LRI Dynamic Model (LRI only included at S1 – reverts to Basic Dynamic)							
	Model development			Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=1197	0.07	63.461	57.5	0.52	-0.024	0.28	S1 *n*=739
S2 *n*=1117	0.18	159.936	66.5	0.653	0.175	0.25	S2 *n*=709
S3 *n*=1138	0.308	295.679	72.8	0.742	0.288	0.22	S3 *n*=719
S4 *n*=921	0.371	298.706	74.5	0.766	0.365	0.21	S4 *n*=508
S5 *n*=665	0.359	207.445	74.3	0.767	0.334	0.2	S5 *n*=361
S6 *n*=345	0.3	87.498	74.8	0.802	0.451	0.18	S6 *n*=235
S7 *n*=187	0.235	36.031	69	0.827	0.423	0.17	S7 *n*=152
S8 *n*=97	0.243	19.523	68	0.816	0.604	0.17	S8 *n*=36

LIT PHQ-9 Progressive Dynamic Model							
	Model development			Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=1197	0.056	50.386	58.1	0.511	-0.052	0.28	S1 *n*=739
S2 *n*=1117	0.187	167.366	67.5	0.664	0.189	0.25	S2 *n*=709
S3 *n*=1138	0.315	303.263	73.9	0.753	0.329	0.22	S3 *n*=719
S4 *n*=921	0.39	316.626	74.6	0.775	0.342	0.21	S4 *n*=508
S5 *n*=665	0.403	237.687	76.4	0.779	0.385	0.21	S5 *n*=361
S6 *n*=345	0.343	101.849	76.2	0.805	0.441	0.19	S6 *n*=235
S7 *n*=187	0.263	40.842	70.6	0.826	0.543	0.17	S7 *n*=152
S8 *n*=97	0.332	27.756	71.1	0.759	0.393	0.2	S8 *n*=36

LIT PHQ-9 Complex Dynamic Model							
	Model development			Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=1197	0.07	63.461	57.5	0.520	-0.024	0.28	S1 *n*=739
S2 *n*=1117	0.187	167.366	67.5	0.664	0.189	0.25	S2 *n*=709
S3 *n*=1138	0.315	303.263	73.9	0.753	0.329	0.22	S3 *n*=719
S4 *n*=921	0.39	316.626	74.6	0.775	0.342	0.21	S4 *n*=508
S5 *n*=665	0.403	237.687	76.4	0.779	0.385	0.21	S5 *n*=361
S6 *n*=345	0.343	101.849	76.2	0.805	0.441	0.19	S6 *n*=235
S7 *n*=187	0.263	40.842	70.6	0.826	0.543	0.17	S7 *n*=152
S8 *n*=97	0.332	27.756	71.1	0.759	0.393	0.2	S8 *n*=36

LIT GAD-7 Models

LIT GAD-7 LRI Profile Model							
Model development				Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=1340	0.038	38.088	58.6	0.575	*0.074	0.27	S1 *n*=857
S2 *n*=1258	0.038	38.088	58.6	0.575	*0.074	0.27	S2 *n*=825
S3 *n*=1288	0.038	38.088	58.6	0.575	*0.074	0.27	S3 *n*=829
S4 *n*=1025	0.027	21.541	56.2	*0.575	0.057	0.27	S4 *n*=590
S5 *n*=715	0.021	*11.492	55.9	**0.564	0.041	0.27	S5 *n*=418
S6 *n*=379	0.012	3.492	56.9	0.537	Constant	0.27	S6 *n*=275
S7 *n*=202	0.01	1.525	56.5	0.540	Constant	0.26	S7 *n*=182
S8 *n*=103	0	.001	55	0.629	constant	0.26	S8 *n*=45

LIT GAD-7 Session Scores only Model							
Model development				Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=1340	0.04	40.001	55.7	0.527	-0.027	0.28	S1 *n*=857
S2 *n*=1258	0.166	166.447	65.3	0.676	0.214	0.24	S2 *n*=825
S3 *n*=1288	0.318	347.405	72.4	0.727	0.280	0.22	S3 *n*=829
S4 *n*=1025	0.321	281.179	72.9	0.724	0.281	0.22	S4 *n*=590
S5 *n*=715	0.358	222.444	73.7	0.757	0.387	0.21	S5 *n*=418
S6 *n*=379	0.26	81.574	70.2	0.764	0.388	0.21	S6 *n*=275
S7 *n*=202	0.263	44.154	73.8	0.826	0.469	0.18	S7 *n*=182
S8 *n*=103	0.153	*12.516	66	*0.733	0.239	0.21	S8 *n*=45

LIT GAD-7 Basic Dynamic Model							
	Model development			Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=1340	0.04	40.001	55.7	0.527	-0.027	0.28	S1 *n*=857
S2 *n*=1258	0.17	170.067	66	0.688	0.246	0.24	S2 *n*=825
S3 *n*=1288	0.33	363.406	72.8	0.741	0.313	0.21	S3 *n*=829
S4 *n*=1025	0.337	297.184	73.9	0.736	0.308	0.22	S4 *n*=590
S5 *n*=715	0.373	233.240	74.1	0.775	0.408	0.2	S5 *n*=418
S6 *n*=379	0.273	86.094	70.7	0.782	0.445	0.2	S6 *n*=275
S7 *n*=202	0.269	45.373	71.8	0.834	0.503	0.17	S7 *n*=182
S8 *n*=103	0.168	13.843	68	*0.750	*0.418	0.21	S8 *n*=45

LIT GAD-7 LRI Dynamic Model							
	Model development			Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=1340	0.056	57.287	58.4	*0.560	0.022	0.27	S1 *n*=857
S2 *n*=1258	0.17	170.067	66	0.688	0.246	0.24	S2 *n*=825
S3 *n*=1288	0.33	363.406	72.8	0.741	0.313	0.21	S3 *n*=829
S4 *n*=1025	0.337	297.184	73.9	0.736	0.308	0.22	S4 *n*=590
S5 *n*=715	0.373	233.240	74.1	0.775	0.408	0.2	S5 *n*=418
S6 *n*=379	0.273	86.094	70.7	0.782	0.445	0.2	S6 *n*=275
S7 *n*=202	0.269	45.373	71.8	0.834	0.503	0.17	S7 *n*=182
S8 *n*=103	0.168	13.843	6830.	*0.750	*0.418	0.21	S8 *n*=45

LIT GAD-7 Progressive Dynamic Model							
	Model development			Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=1340	0.04	40.001	55.7	0.527	-0.027	0.28	S1 *n*=857
S2 *n*=1258	0.181	181.938	66.5	0.697	0.259	0.24	S2 *n*=825
S3 *n*=1288	0.347	384.234	73.6	0.749	0.315	0.22	S3 *n*=829
S4 *n*=1025	0.375	336.299	74.9	0.747	0.340	0.22	S4 *n*=590
S5 *n*=715	0.418	267.751	75.9	0.789	0.438	0.21	S5 *n*=418
S6 *n*=379	0.345	112.437	75.5	0.779	0.389	0.22	S6 *n*=275
S7 *n*=202	0.336	58.405	73.3	0.832	0.507	0.18	S7 *n*=182
S8 *n*=103	0.258	22.052	72.8	*0.756	0.252	0.2	S8 *n*=45

LIT GAD-7 Complex Dynamic Model							
	Model development			Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=1340	0.056	57.287	58.4	*0.560	0.022	0.27	S1 *n*=857
S2 *n*=1258	0.181	181.938	66.5	0.697	0.259	0.24	S2 *n*=825
S3 *n*=1288	0.347	384.234	73.6	0.749	0.315	0.22	S3 *n*=829
S4 *n*=1025	0.375	336.299	74.9	0.747	0.340	0.22	S4 *n*=590
S5 *n*=715	0.418	267.751	75.9	0.789	0.438	0.21	S5 *n*=418
S6 *n*=379	0.345	112.437	75.5	0.779	0.389	0.22	S6 *n*=275
S7 *n*=202	0.336	58.405	73.3	0.832	0.507	0.18	S7 *n*=182
S8 *n*=103	0.258	22.052	72.8	*0.756	0.252	0.2	S8 *n*=45

HIT PHQ-9 Models

HIT PHQ-9 LRI Profile Model							
Model development				Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=860	0.043	28.392	57.7	0.598	0.183	0.24	S1 *n*=1040
S2 *n*=811	0.043	28.392	57.7	0.598	0.183	0.24	S2 *n*=1000
S3 *n*=835	0.043	28.392	57.7	0.598	0.183	0.24	S3 *n*=997
S4 *n*=764	0.042	25.099	58.2	0.596	0.182	0.24	S4 *n*=853
S5 *n*=705	0.034	18.528	58.7	0.590	0.176	0.24	S5 *n*=716
S6 *n*=636	0.035	17.273	59.6	0.592	0.179	0.23	S6 *n*=589
S7 *n*=591	0.033	14.886	59.7	*0.593	0.180	0.23	S7 *n*=479
S8 *n*=527	0.026	*10.674	59.6	*0.605	*0.177	0.22	S8 *n*=356
S9 *n*=473	0.027	*9.948	59.9	**0.589	*0.171	0.23	S9 *n*=283
S10 *n*=424	0.022	**7.126	60.1	**0.588	*0.174	0.23	S10 *n*=220
S11 *n*=362	0.026	**7.187	60.6	0.569	**0.150	0.23	S11 *n*=176
S12 *n*=302	0.026	**6.190	29.5	0.575	0.152	0.23	S12 *n*=125

HIT PHQ-9 Session Scores only Model							
Model development				Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=860	0.021	13.933	57.7	0.585	0.142	0.25	S1 *n*=1040
S2 *n*=811	0.109	68.928	59.8	0.663	0.213	0.23	S2 *n*=1000
S3 *n*=835	0.164	109.381	65.3	0.726	0.290	0.21	S3 *n*=997
S4 *n*=764	0.182	112.019	66.5	0.755	0.343	0.2	S4 *n*=853
S5 *n*=705	0.237	137.566	67.4	0.764	0.393	0.19	S5 *n*=716
S6 *n*=636	0.232	120.616	67.9	0.777	0.447	0.18	S6 *n*=589
S7 *n*=591	0.254	123.982	67.3	0.802	0.444	0.17	S7 *n*=479
S8 *n*=527	0.28	122.768	71.2	0.820	0.522	0.16	S8 *n*=356
S9 *n*=473	0.295	117.200	71	0.800	0.389	0.18	S9 *n*=283
S10 *n*=424	0.325	116.785	73.1	0.829	0.440	0.17	S10 *n*=220
S11 *n*=362	0.337	104.181	72.1	0.820	0.495	0.17	S11 *n*=176
S12 *n*=302	0.341	87.465	73.2	0.830	0.461	0.17	S12 *n*=125

HIT PHQ-9 Basic Dynamic Model							
	Model development			Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=860	0.021	13.933	57.7	0.585	0.142	0.24	S1 *n*=1040
S2 *n*=811	0.123	78.273	61.7	0.669	0.221	0.23	S2 *n*=1000
S3 *n*=835	0.175	117.364	65.1	0.735	0.301	0.2	S3 *n*=997
S4 *n*=764	0.194	120.104	66.4	0.766	0.389	0.19	S4 *n*=853
S5 *n*=705	0.258	151.289	67.4	0.775	0.415	0.18	S5 *n*=716
S6 *n*=636	0.254	133.508	69.8	0.781	0.409	0.18	S6 *n*=589
S7 *n*=591	0.275	135.626	68.4	0.806	0.489	0.17	S7 *n*=479
S8 *n*=527	0.298	131.822	71.3	0.826	0.524	0.16	S8 *n*=356
S9 *n*=473	0.322	129.239	71.9	0.799	0.395	0.18	S9 *n*=283
S10 *n*=424	0.36	131.447	74.1	0.821	0.405	0.17	S10 *n*=220
S11 *n*=362	0.372	116.989	74	0.804	0.435	0.18	S11 *n*=176
S12 *n*=302	0.369	95.932	77.2	0.809	0.470	0.18	S12 *n*=125

HIT PHQ-9 LRI Dynamic Model (includes LRI up to S6 then reverts to basic dynamic)							
	Model development			Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=860	0.047	30.659	57.8	0.614	0.192	0.24	S1 *n*=1040
S2 *n*=811	0.142	91.105	64.1	0.678	0.217	0.22	S2 *n*=1000
S3 *n*=835	0.187	125.958	65.3	0.738	0.304	0.2	S3 *n*=997
S4 *n*=764	0.21	130.514	67.4	0.765	0.404	0.19	S4 *n*=853
S5 *n*=705	0.267	157.050	68.4	0.777	0.424	0.18	S5 *n*=716
S6 *n*=636	0.27	143.135	71.2	0.778	0.413	0.18	S6 *n*=589
S7 *n*=591	0.275	135.626	68.4	0.806	0.489	0.17	S7 *n*=479
S8 *n*=527	0.298	131.822	71.3	0.826	0.524	0.16	S8 *n*=356
S9 *n*=473	0.322	129.239	71.9	0.799	0.395	0.18	S9 *n*=283
S10 *n*=424	0.36	131.447	74.1	0.821	0.405	0.17	S10 *n*=220
S11 *n*=362	0.372	116.989	74	0.804	0.435	0.18	S11 *n*=176
S12 *n*=302	0.369	95.932	77.2	0.809	0.470	0.18	S12 *n*=125

HIT PHQ-9 Progressive Dynamic Model							
	Model development			Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=860	0.021	13.933	57.7	0.585	0.142	0.24	S1 *n*=1040
S2 *n*=811	0.123	78.273	61.7	0.669	0.221	0.23	S2 *n*=1000
S3 *n*=835	0.181	121.600	66.7	0.737	0.308	0.2	S3 *n*=997
S4 *n*=764	0.201	124.906	66.6	0.769	0.384	0.19	S4 *n*=853
S5 *n*=705	0.265	155.253	68.5	0.780	0.423	0.18	S5 *n*=716
S6 *n*=636	0.274	145.275	69.2	0.788	0.416	0.17	S6 *n*=589
S7 *n*=591	0.29	143.611	68.4	0.814	0.470	0.16	S7 *n*=479
S8 *n*=527	0.325	145.544	72.7	0.826	0.530	0.16	S8 *n*=356
S9 *n*=473	0.336	136.019	73.6	0.811	0.417	0.17	S9 *n*=283
S10 *n*=424	0.371	135.951	74.1	0.830	0.445	0.16	S10 *n*=220
S11 *n*=362	0.385	121.654	74.6	0.816	0.386	0.17	S11 *n*=176
S12 *n*=302	0.387	101.404	75.2	0.824	0.489	0.17	S12 *n*=125

HIT PHQ-9 Complex Dynamic Model							
	Model development			Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=860	0.047	30.659	57.8	0.614	0.192	0.24	S1 *n*=1040
S2 *n*=811	0.142	91.105	64.1	0.678	0.217	0.22	S2 *n*=1000
S3 *n*=835	0.194	131.081	66.6	0.739	0.341	0.2	S3 *n*=997
S4 *n*=764	0.219	136.696	67.4	0.769	0.411	0.19	S4 *n*=853
S5 *n*=705	0.275	161.838	68.8	0.783	0.444	0.18	S5 *n*=716
S6 *n*=636	0.283	150.915	70.8	0.783	0.425	0.18	S6 *n*=589
S7 *n*=591	0.29	143.611	68.4	0.814	0.470	0.16	S7 *n*=479
S8 *n*=527	0.325	145.544	72.7	0.826	0.530	0.16	S8 *n*=356
S9 *n*=473	0.336	136.019	73.6	0.811	0.417	0.17	S9 *n*=283
S10 *n*=424	0.371	135.951	74.1	0.830	0.445	0.16	S10 *n*=220
S11 *n*=362	0.385	121.654	74.6	0.816	0.386	0.17	S11 *n*=176
S12 *n*=302	0.387	101.404	75.2	0.824	0.489	0.17	S12 *n*=125

HIT GAD-7 Models

HIT GAD-7 LRI Profile Model							
Model development				Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=976	0.055	41.191	58.1	0.604	0.171	0.24	S1 *n*=1107
S2 *n*=924	0.055	41.191	58.1	0.604	0.171	0.24	S2 *n*=1069
S3 *n*=945	0.055	41.191	58.1	0.604	0.171	0.24	S3 *n*=1063
S4 *n*=862	0.056	38.361	59.6	0.603	0.171	0.24	S4 *n*=903
S5 *n*=801	0.05	31.160	60.2	0.608	0.173	0.24	S5 *n*=760
S6 *n*=731	0.05	28.618	61.1	0.597	0.159	0.24	S6 *n*=618
S7 *n*=678	0.044	23.200	61.4	0.597	*0.146	0.23	S7 *n*=509
S8 *n*=607	0.041	19.206	61.9	*0.591	**0.111	0.24	S8 *n*=383
S9 *n*=546	0.049	20.642	62.6	*0.597	**0.131	0.23	S9 *n*=305
S10 *n*=482	0.033	*12.449	61.7	*0.607	**0.151	0.23	S10 *n*=238
S11 *n*=419	0.042	13.510	62.6	**0.592	0.138	0.24	S11 *n*=190
S12 *n*=345	0.038	*10.238	62.5	0.572	0.105	0.24	S12 *n*=133

HIT GAD-7 Session Scores only Model							
Model development				Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=976	0.019	13.845	58.1	0.577	0.142	0.25	S1 *n*=1107
S2 *n*=924	0.098	70.703	62.1	0.667	0.239	0.23	S2 *n*=1069
S3 *n*=945	0.187	142.881	65.1	0.723	0.319	0.21	S3 *n*=1063
S4 *n*=862	0.176	122.070	64.3	0.730	0.324	0.21	S4 *n*=903
S5 *n*=801	0.213	138.699	67.2	0.758	0.374	0.19	S5 *n*=760
S6 *n*=731	0.243	146.159	68.4	0.757	0.412	0.2	S6 *n*=618
S7 *n*=678	0.274	154.473	71.4	0.799	0.460	0.18	S7 *n*=509
S8 *n*=607	0.306	156.176	72.5	0.772	0.384	0.19	S8 *n*=383
S9 *n*=546	0.332	154.199	73.4	0.786	0.391	0.19	S9 *n*=305
S10 *n*=482	0.321	130.902	73.4	0.785	0.388	0.19	S10 *n*=238
S11 *n*=419	0.387	141.308	75.4	0.798	0.446	0.19	S11 *n*=190
S12 *n*=345	0.394	118.678	75.7	0.797	0.396	0.19	S12 *n*=133

HIT GAD-7 Basic Dynamic Model							
	Model development			Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=976	0.019	13.845	58.1	0.577	0.142	0.25	S1 *n*=1107
S2 *n*=924	0.107	77.302	62	0.672	0.240	0.23	S2 *n*=1069
S3 *n*=945	0.201	154.508	66.9	0.732	0.286	0.21	S3 *n*=1063
S4 *n*=862	0.184	128.026	63.6	0.738	0.371	0.2	S4 *n*=903
S5 *n*=801	0.223	145.871	67.4	0.767	0.389	0.19	S5 *n*=760
S6 *n*=731	0.258	156.122	69.2	0.764	0.376	0.19	S6 *n*=618
S7 *n*=678	0.281	159.203	70.6	0.801	0.442	0.18	S7 *n*=509
S8 *n*=607	0.308	157.114	72.3	0.774	0.388	0.19	S8 *n*=383
S9 *n*=546	0.335	156.152	73.1	0.791	0.371	0.18	S9 *n*=305
S10 *n*=482	0.326	133.097	74.1	0.784	0.386	0.19	S10 *n*=238
S11 *n*=419	0.402	147.978	76.8	0.800	0.422	0.18	S11 *n*=190
S12 *n*=345	0.423	129.233	75.9	0.791	0.379	0.19	S12 *n*=133

HIT GAD-7 LRI Dynamic Model (LRI included at all sessions)							
	Model development			Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=976	0.059	44.458	59.5	0.620	0.174	0.24	S1 *n*=1107
S2 *n*=924	0.134	97.774	64.4	0.686	0.260	0.23	S2 *n*=1069
S3 *n*=945	0.222	172.340	66.9	0.734	0.315	0.21	S3 *n*=1063
S4 *n*=862	0.214	150.679	66.6	0.740	0.363	0.2	S4 *n*=903
S5 *n*=801	0.242	159.518	68	0.775	0.385	0.19	S5 *n*=760
S6 *n*=731	0.276	*168.469	70.9	0.765	0.366	0.19	S6 *n*=618
S7 *n*=678	0.291	**165.170	71.2	0.804	0.486	0.18	S7 *n*=509
S8 *n*=607	0.321	*165.206	74.1	0.773	0.382	0.19	S8 *n*=383
S9 *n*=546	0.35	**164.172	74.4	0.789	0.404	0.19	S9 *n*=305
S10 *n*=482	0.34	**139.740	74.9	0.787	0.373	0.19	S10 *n*=238
S11 *n*=419	0.416	**154.468	77.6	0.804	0.456	0.19	S11 *n*=190
S12 *n*=345	0.441	**136.112	76.5	0.789	0.422	0.2	S12 *n*=133

HIT GAD-7 Progressive Dynamic Model							
	Model development			Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=976	0.019	13.845	58.1	0.577	0.142	0.24	S1 *n*=1107
S2 *n*=924	0.107	77.302	62	0.672	0.240	0.23	S2 *n*=1069
S3 *n*=945	0.201	154.508	66.9	0.732	0.286	0.21	S3 *n*=1063
S4 *n*=862	0.212	148.701	65.9	0.746	0.381	0.2	S4 *n*=903
S5 *n*=801	0.254	168.811	69	0.765	0.386	0.19	S5 *n*=760
S6 *n*=731	0.285	174.469	71	0.767	0.396	0.19	S6 *n*=618
S7 *n*=678	0.303	173.236	73.2	0.805	0.466	0.18	S7 *n*=509
S8 *n*=607	0.345	179.023	73.3	0.777	0.416	0.19	S8 *n*=383
S9 *n*=546	0.365	172.392	72.9	0.802	0.438	0.18	S9 *n*=305
S10 *n*=482	0.374	156.265	74.9	0.802	0.388	0.18	S10 *n*=238
S11 *n*=419	0.42	156.047	77.1	0.822	0.469	0.17	S11 *n*=190
S12 *n*=345	0.449	138.815	75.9	0.812	0.487	0.18	S12 *n*=133

HIT GAD-7 Complex Dynamic Model							
	Model development			Cross-validation			
	Variance N_2	X_2	% correct	AUC	Kappa	Brier	
S1 *n*=976	0.059	44.458	59.5	0.620	0.174	0.24	S1 *n*=1107
S2 *n*=924	0.134	97.774	64.4	0.686	0.260	0.23	S2 *n*=1069
S3 *n*=945	0.222	172.340	66.9	0.734	0.315	0.21	S3 *n*=1063
S4 *n*=862	0.237	168.450	67.2	0.746	0.348	0.2	S4 *n*=903
S5 *n*=801	0.269	179.680	69.5	0.772	0.385	0.19	S5 *n*=760
S6 *n*=731	0.298	183.874	71.8	0.767	0.402	0.19	S6 *n*=618
S7 *n*=678	0.303	173.236	73.2	0.805	0.466	0.18	S7 *n*=509
S8 *n*=607	0.345	179.023	73.3	0.777	0.416	0.19	S8 *n*=383
S9 *n*=546	0.365	172.392	72.9	0.802	0.438	0.18	S9 *n*=305
S10 *n*=482	0.374	156.265	74.9	0.802	0.388	0.18	S10 *n*=238
S11 *n*=419	0.42	156.047	77.1	0.822	0.469	0.17	S11 *n*=190
S12 *n*=345	0.449	138.815	75.9	0.812	0.487	0.18	S12 *n*=133

Appendix B

Proof of Ethical Approval

NHS
Health Research Authority

Dr Jaime Delgadillo Clinical Psychology Unit Floor F, Cathedral Court Sheffield S1 2LT

16 January 2018 Dear Dr Delgadillo,

Study title: The Development of a Dynamic Progress System to Guide Psychological Treatment in Primary Care.

IRAS project ID: REC reference: Sponsor

Email: hra.approval@nhs.net

Letter of HRA Approval

Development of a dynamic progress feedback system to guide psychological treatment in primary care.
233799
18/WM/0012

Cumbria Partnership NHS Foundation Trust

I am pleased to confirm that HRA Approval has been given for the above referenced study, on the basis described in the application form, protocol, supporting documentation and any clarifications noted in this letter.

Participation of NHS Organisations in England

The sponsor should now provide a copy of this letter to all participating NHS organisations in England.

Appendix B provides important information for sponsors and participating NHS organisations in England for arranging and confirming capacity and capability. Please read Appendix B carefully, in particular the following sections:

- • Participating NHS organisations in England – this clarifies the types of participating organisations in the study and whether or not all organisations will be undertaking the same activities
- • Confirmation of capacity and capability - this confirms whether or not each type of participating NHS organisation in England is expected to give formal

confirmation of capacity and capability. Where formal confirmation is not expected, the section also provides details on the time limit given to participating organisations to opt out of the study, or request additional time, before their participation is assumed.

- • Allocation of responsibilities and rights are agreed and documented (4.1 of HRA assessment criteria) - this provides detail on the form of agreement to be used in the study to confirm capacity and capability, where applicable.

  Further information on funding, HR processes, and compliance with HRA criteria and standards is also provided.

  It is critical that you involve both the research management function (e.g. R&D office) supporting each organisation and the local research team (where there is one) in setting up your study. Contact details

---

and further information about working with the research management function for each organisation can be accessed from the HRA website.

Appendices

The HRA Approval letter contains the following appendices:

- • A – List of documents reviewed during HRA assessment
- • B – Summary of HRA assessment After HRA Approval

  The document “After Ethical Review – guidance for sponsors and investigators”, issued with your REC favourable opinion, gives detailed guidance on reporting expectations for studies, including:

- • Registration of research
- • Notifying amendments
- • Notifying the end of the study

  The HRA website also provides guidance on these topics, and is updated in the light of changes in reporting expectations or procedures.

  In addition to the guidance in the above, please note the following:

- • HRA Approval applies for the duration of your REC favourable opinion, unless otherwise notified in writing by the HRA.
- • Substantial amendments should be submitted directly to the Research Ethics Committee, as detailed in the After Ethical Review document. Non-substantial

amendments should be submitted for review by the HRA using the form provided on the HRA website, and emailed to hra.amendments@nhs.net.

- • The HRA will categorise amendments (substantial and non-substantial) and issue confirmation of continued HRA Approval. Further details can be found on the HRA website.

Scope

HRA Approval provides an approval for research involving patients or staff in NHS organisations in England.

If your study involves NHS organisations in other countries in the UK, please contact the relevant national coordinating functions for support and advice. Further information can be found through IRAS.

If there are participating non-NHS organisations, local agreement should be obtained in accordance with the procedures of the local participating non-NHS organisation.

User Feedback

The Health Research Authority is continually striving to provide a high quality service to all applicants and sponsors. You are invited to give your view of the service you have received and the application procedure. If you wish to make your views known please use the feedback form available on the HRA website.

IRAS project ID	233799

Your IRAS project ID is 233799. Please quote this on all correspondence. Yours sincerely

Kevin Ahmed Assessor

Telephone: 0207 104 8171 Email: hra.approval@nhs.net

Appendix C

Sample Size Calculation Table (taken from Hsieh,1989).

Table III. Sample size required for univariate logistic regression having an overall event proportion P and an odds ratio r at one standard deviation above the mean of the covariate when $\alpha = 5$ per cent (one-tailed) and $1-\beta = 90$ per cent

	Odds ratio r															
P	0·6	0·7	0·8	0·9	1·1	1·2	1·3	1·4	1·5	1·6	1·7	1·8	1·9	2·0	2·5	3·0
0·01	3192	6706	17383	78551	96029	26120	12529	7554	5154	3797	2948	2377	1972	1674	917	627
0·02	1640	3430	8873	40056	48966	13327	6398	3863	2639	1948	1516	1225	1020	869	488	349
0·03	1123	2338	6036	27225	33279	9063	4355	2632	1801	1332	1038	842	702	600	345	256
0·04	864	1792	4618	20809	25435	6930	3333	2017	1382	1024	800	650	544	466	274	210
0·05	709	1465	3767	16959	20729	5651	2720	1648	1131	839	657	534	448	385	231	182
0·06	605	1246	3199	14393	17591	4798	2311	1402	963	715	561	458	385	332	202	163
0·07	532	1090	2794	12560	15350	4189	2019	1226	843	627	493	403	340	293	182	150
0·08	476	973	2490	11185	13670	3732	1800	1094	753	561	442	362	306	265	167	140
0·09	433	882	2254	10116	12362	3377	1630	991	683	510	402	330	279	242	155	132
0·10	398	810	2065	9260	11317	3092	1494	909	628	469	370	304	258	224	145	126
0·12	347	700	1781	7977	9748	2666	1289	786	544	407	322	266	226	197	131	117
0·14	310	622	1578	7061	8627	2361	1143	698	484	363	288	238	203	178	121	110
0·16	282	564	1426	6373	7787	2133	1034	632	439	330	263	218	186	164	113	105
0·18	261	518	1308	5839	7133	1955	949	581	404	305	243	202	173	153	107	101
0·20	243	482	1214	5411	6610	1813	881	540	376	284	227	189	163	144	102	98
0·25	212	417	1043	4641	5669	1557	758	466	326	247	198	166	144	128	94	93
0·30	192	373	930	4128	5042	1387	676	417	292	222	179	151	131	117	88	89
0·35	177	342	849	3761	4593	1265	618	382	268	205	166	140	122	109	84	86
0·40	166	318	788	3486	4257	1173	574	355	250	192	155	131	115	103	81	84
0·45	157	300	741	3272	3996	1102	540	335	236	181	147	125	110	99	78	83
0·50	150	286	703	3101	3787	1045	513	319	225	173	141	120	105	95	76	81

Note: To obtain sample sizes for multiple logistic regression, divide the number from the table by a factor of $1-\rho^2$, where ρ is the multiple correlation coefficient relating the specific covariate to the remaining covariates.

Table IV. Sample size required for univariate logistic regression having an overall event proportion P and an odds ratio r at one standard deviation above the mean of the covariate when $\alpha = 5$ per cent (one-tailed) and $1-\beta = 95$ per cent

	Odds ratio r															
P	0·6	0·7	0·8	0·9	1·1	1·2	1·3	1·4	1·5	1·6	1·7	1·8	1·9	2·0	2·5	3·0
0·01	4001	8439	21927	99209	121290	32967	15795	9511	6478	4765	3692	2971	2461	2084	1130	764
0·02	2055	4316	11192	50591	61847	16820	8066	4863	3318	2445	1898	1532	1272	1081	601	425
0·03	1407	2942	7614	34384	42033	11438	5490	3314	2264	1671	1301	1052	876	747	425	312
0·04	1083	2255	5825	26281	32126	8747	4202	2539	1737	1284	1002	812	678	580	337	255
0·05	888	1843	4751	21419	26182	7132	3429	2074	1421	1052	822	668	559	480	284	221
0·06	759	1568	4036	18178	22219	6056	2914	1764	1210	898	703	572	480	413	249	199
0·07	666	1372	3525	15863	19389	5287	2546	1543	1060	787	617	504	424	365	224	183
0·08	597	1225	3141	14127	17266	4710	2270	1377	947	704	553	452	381	329	205	170
0·09	543	1110	2843	12776	15614	4262	2055	1248	859	640	503	412	348	301	190	161
0·10	499	1019	2604	11696	14293	3903	1883	1145	789	588	464	380	322	279	179	153
0·12	435	881	2247	10075	12312	3365	1626	990	684	511	404	332	282	246	161	142
0·14	388	783	1991	8918	10897	2980	1442	879	608	456	361	298	254	222	148	134
0·16	353	710	1799	8049	9835	2692	1304	796	552	414	329	272	233	204	139	128
0·18	326	652	1650	7374	9010	2468	1196	732	508	382	304	252	216	190	132	123
0·20	305	607	1531	6834	8349	2288	1110	680	473	356	284	236	203	179	126	120
0·25	266	524	1316	5861	7160	1965	956	587	410	310	248	207	179	159	115	113
0·30	240	469	1173	5213	6368	1750	853	525	367	279	224	188	163	145	108	108
0·35	222	430	1071	4750	5802	1596	779	481	337	257	207	175	152	136	103	105
0·40	208	401	994	4403	5377	1481	724	448	315	240	194	164	143	129	99	103
0·45	197	378	935	4133	5047	1391	681	422	297	228	185	156	137	123	96	101
0·50	188	359	887	3917	4783	1319	647	401	283	217	177	150	132	119	94	99

Note: To obtain sample sizes for multiple logistic regression, divide the number from the table by a factor of $1-\rho^2$, where ρ is the multiple correlation coefficient relating the specific covariate to the remaining covariates.

Appendix D

Outcome Measures

Generalized Anxiety Disorder 7-item (GAD-7) scale

Over the last 2 weeks, how often have you been bothered by the following problems?	Not at all sure	Several days	Over half the days	Nearly every day
1. Feeling nervous, anxious, or on edge	0	1	2	3
2. Not being able to stop or control worrying	0	1	2	3
3. Worrying too much about different things	0	1	2	3
4. Trouble relaxing	0	1	2	3
5. Being so restless that it's hard to sit still	0	1	2	3
6. Becoming easily annoyed or irritable	0	1	2	3
7. Feeling afraid as if something awful might happen	0	1	2	3
Add the score for each column	+	+	+	
Total Score *(add your column scores)* =				

If you checked off any problems, how difficult have these made it for you to do your work, take care of things at home, or get along with other people?

Not difficult at all _____
Somewhat difficult _____
Very difficult _____
Extremely difficult _____

PATIENT HEALTH QUESTIONNAIRE-9 (PHQ-9)

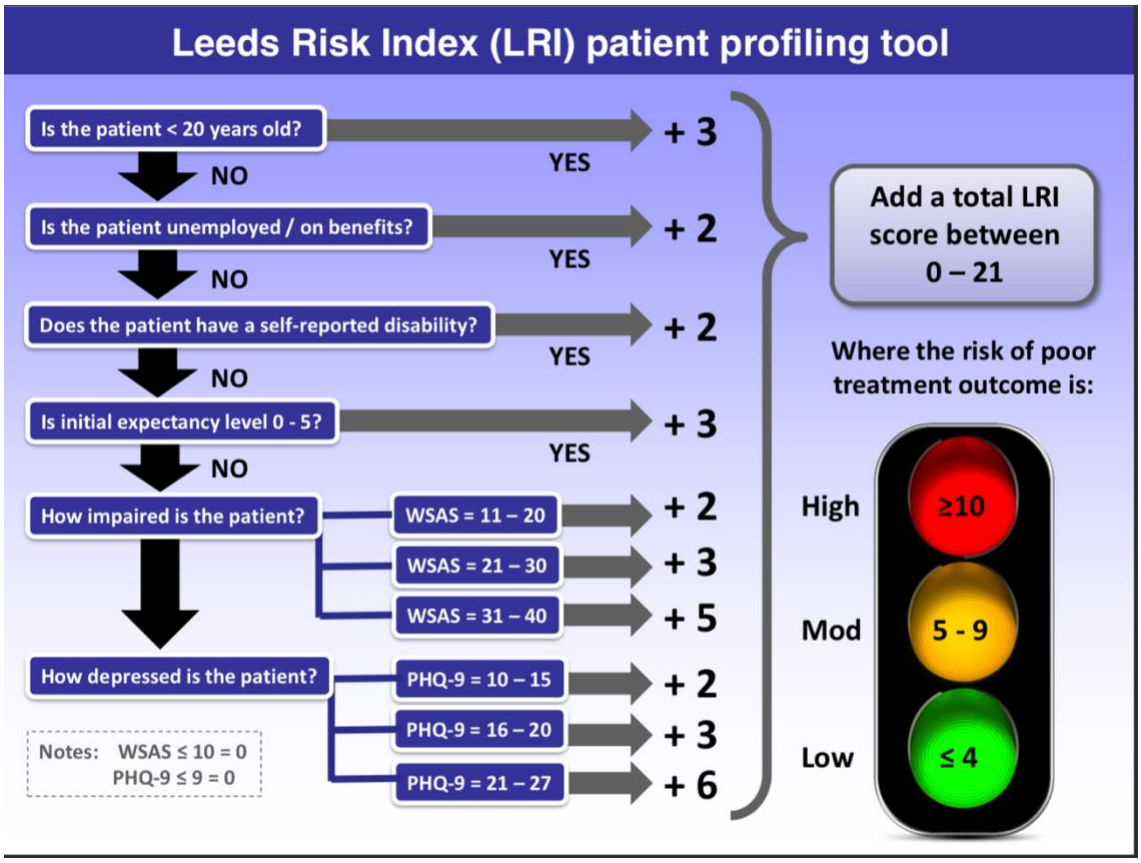
Over the last 2 weeks, how often have you been bothered by any of the following problems? *(Use "✔" to indicate your answer)*	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself — or that you are a failure or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed? Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
9. Thoughts that you would be better off dead or of hurting yourself in some way	0	1	2	3

FOR OFFICE CODING 0 + _____ + _____ + _____

=Total Score: _____

If you checked off any problems, how difficult have these problems made it for you to do your work, take care of things at home, or get along with other people?

Not difficult at all	Somewhat difficult	Very difficult	Extremely difficult
☐	☐	☐	☐


Leeds Risk Index (LRI) patient profiling tool
Is the patient < 20 years old?
YES
+ 3
NO
Is the patient unemployed / on benefits?
YES
+ 2
NO
Does the patient have a self-reported disability?
YES
+ 2
NO
Is initial expectancy level 0 - 5?
YES
+ 3
NO
How impaired is the patient?
WSAS = 11 – 20
+ 2
WSAS = 21 – 30
+ 3
WSAS = 31 – 40
+ 5
How depressed is the patient?
PHQ-9 = 10 – 15
+ 2
PHQ-9 = 16 – 20
+ 3
PHQ-9 = 21 – 27
+ 6
Notes: WSAS ≤ 10 = 0
PHQ-9 ≤ 9 = 0
Add a total LRI score between 0 – 21
Where the risk of poor treatment outcome is:
High
≥10
Mod
5 - 9
Low
≤ 4

Appendix E

IAPT Consent Form

Leeds Community Healthcare

INFORMATION ABOUT STORING AND SHARING

YOUR CONFIDENTIAL INFORMATION

This leaflet gives details about the information we need to ensure that we provide you with a high quality service. It explains what happens to the information you provide and how you will be involved in sharing it. This leaflet gives you answers to commonly asked questions about how we store your confidential information, your right to access this information and our usual NHS practice of confidentiality.

If you have questions or concerns you can telephone us during office hours on the same number you used to make an appointment. It is important to us that you are happy with the arrangements we have made for your care, so please feel comfortable calling us if you are unsure. If after speaking with us you are still not happy you can contact PALS on 0800 0525790 who will be able to help you further.

What kind of information do you keep?

We keep contact information for you and others involved in your care, information about your background, assessments, results of tests and questionnaires, our plans for your future care, details of the care we give you and correspondence related to your care. It is important that you tell us within one week if you change your details, telephone numbers or address because we will continue to use the address and telephone numbers you have given us until you tell us they have changed.

How do you store information about my care?

We keep information about your care in paper records and on a specialist and secure computer system.

What are each of these used for?

The paper records contain notes and copies of documents related to your care. Our computer systems contain electronic records of your care. These systems are used by staff to plan and monitor the quality of your care, to conduct audit and research in order to continually improve the quality of the services that we offer, and to plan future services.

Can I see my records?

Yes, we are happy to provide you with a copy of your records and you will need to write to us to request these (there may be a standard copying fee) or if appropriate we can meet with you to read and discuss your notes together.

Who will know about my care?

You have control over who else is involved in your care and this service observes strict NHS standards of confidentiality. The only time we will inform others without your permission is if we are very concerned for your immediate safety, for the safety of someone else, or if a British Court orders the release of your records. We will try to contact you first if this happens and do our best to help you.

We work in partnership with three voluntary sector organisations in Leeds, Community Links, Leeds Counselling and Touchstone. After discussing with you, you may be offered an appointment with one of these organisations and with your permission information will be shared. All organisations adhere to strict NHS standards of confidentiality.

We will write to your GP about your care; this is usual in the NHS as your GP is the main person who organises your care.

How does the service use the questionnaires and other information to improve my care?

After you have completed the questionnaires we enter your results into our secure computer system. We use the results to plan your care. You can ask for a print out of your results from your therapist to show how much you have improved.

How is the information used to improve the service offered?

After we have removed all your details from the results, we collect together all the results from all the patients. This means that someone who looks at the data cannot tell who gave the replies (the data is anonymous) and it is impossible to identify any individual patient. We use these results to look for ways to improve the service we offer through audit and research. We also provide this anonymous data to organisations that pay for the service we offer and share what we have learned with other health professionals. If you wish to find out further details about how anonymous information is used in audit, research and reporting, or if you wish to withdraw your consent to share your information for these purposes, please contact us on the number provided on the front page of this leaflet.

How can I help?

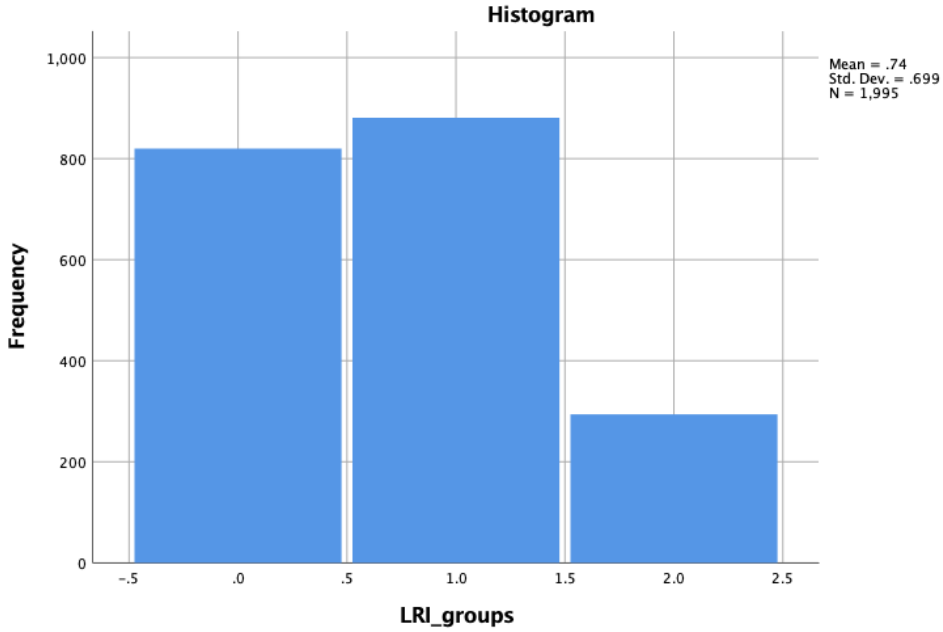
As part of your treatment you will be asked to complete some questionnaires. These questionnaires are not compulsory; however, they are an important part of your treatment and we use them to tailor your care to your individual needs. In addition, without these results it is more difficult to assess your improvement and we cannot show how we are helping people.

If you have further questions please ask to speak with a member of the team.

Appendix F

LRI Groups in Dataset One Model Development

LRI Low Intensity



LRI High Intensity